THE COMPARABILITY BETWEEN MODULAR AND NON-MODULAR EXAMINATIONS AT GCE ADVANCED LEVEL

Elizabeth A Gray

PhD

INSTITUTE OF EDUCATION
UNIVERSITY OF LONDON

ABSTRACT

The prime concern of this thesis is the comparability of two types of assessment now prevalent in Advanced level GCE examinations. The more conventional linear scheme assesses all candidates terminally, and the only way to improve the grade awarded is to re-take the whole examination. In contrast, the relatively new modular schemes of assessment include testing opportunities throughout the course of study. This not only has formative effects but allows quantifiable improvements in syllabus results through the medium of the resit option. There are obvious differences between the two schemes, but this does not necessarily imply that they are not comparable in their grading standards. It is this standard which the thesis attempts to address by considering the different variabilities of each of the schemes, and how these might impinge upon the outcomes of the grading process as evidenced in the final grade distributions. A key issue is that of legitimate and illegitimate variabilities - the former perhaps allowing an improvement in performance while maintaining grading standards; the latter possibly affecting the grading standard because its effect was not fully taken into account in the awarding process.

By looking at a linear and modular syllabus in mathematics, the differences between the two are investigated, and although not fully generalisable, it is clear that many of the worries which were advanced when modular schemes were first introduced are groundless. Most candidates are seen to use the testing flexibility to their advantage, but there is little evidence of over-testing. Perhaps the major finding is a negative one - that there is no clear evidence for any difference in grading standards between modular and linear schemes of assessment, although there are variabilities which go some way to explaining what appears to be enhanced performances by some of the weaker modular candidates.

INDEX

List of Tables

Table:

List of Figures

Figure:

ACKNOWLEDGEMENTS

First of all I must acknowledge my debt to the Oxford and Cambridge Schools Examination Board, in the person of its Secretary Mr H.F. King, who was not only prepared to finance this Ph.D. thesis, but also allowed me to use the data on which this research is based: and to UCLES who took over responsibility for my fees when the two Boards merged. For

a woman must have money and a room of her own if she is to write

Virginia Woolf 1928

I am also grateful to the Joint Forum for the GCSE and GCE who allowed me to use the data from the 16/18 database.

I must also thank my supervisor, Professor Harvey Goldstein, without whose gentle persistence and expertise the completion of this thesis would not have been possible. To Dr Helen Patrick thanks are due not only for her encouragement, but also for reading parts of this thesis and offering intelligent and helpful suggestions. Also to Mr Alastair Pollitt who allowed me time to do my thinking and writing. Finally I have to thank my husband and children who let me get on with it, in a room of my own!

To all these, and more, I give my thanks, but any mistakes within this thesis I, of course, acknowledge as entirely my own.

CHAPTER 1

The Hunt for Comparability

At first glance, the GCE A level examination has remained essentially unchanged since it was introduced in 1951. It is still primarily a terminal examination (though this primacy is rapidly disappearing) which has the dual purpose of summarising attainment in specified subjects and judging the fitness of 17 and 18 year olds for higher education. Syllabuses have been regularly updated and are all intended to provide the basis for a two year course of study, post O level/GCSE. However, although it is seen as being the "gold standard" A level has had to undergo some re-thinking of late. The developments in GCSE and introduction of the National Curriculum have meant that preparation for A level courses has changed. For example, the double award science GCSE has led to some misgivings on behalf of candidates who wish to study a single science at A level, and this has led in turn to the introduction of "Science" A levels, often as part of a modular scheme. The increase in numbers now studying post-16 has also meant that provision must be made for a wider range of skills. Indeed, the Dearing report (1993) suggests that the target is for 50% of the 17/18 age cohort to be obtaining two A levels, or their vocational equivalent, by the year 2000. This would be a natural consequence if the number of candidates gaining 4 or more GCSEs (or the vocational equivalent) at grade C or above increases from 47% to the 80% target figure. The introduction of agreed syllabus cores in 1983 for mainstream subjects (recently updated for gradual incorporation from 1996 onwards) has meant that some parity of content is guaranteed between examinations set by different boards in those subjects. The most innovative of recent changes has been the successful introduction of modular courses. These bring with them a new philosophy enabling resits within the course of study and assessment arrangements which must be continuous from year to year. Current sources of contention within the modular world, the imposition of a fixed minimum percentage of the assessment having to be achieved in a terminal examination and the debate over a better form of uniform marks policy will be resolved. By 1996, most A level boards were certificating A and AS levels for modular schemes in many main-stream subjects - mathematics, English, science, modern languages and humanities. Other subjects are being re-structured so that they can be easily "modularised" if required. A very gradual change is taking place, though it is gathering pace with the introduction of new syllabus cores and syllabuses. It is thus even more important if standards are to be maintained that comparability is not jeopardised by this "quiet reformation".

It is instructive to note the words of UCAS (1994) on the subject of modular syllabuses:

It should be clearly understood that modular syllabuses are no easy option, as all modules are assessed to full GCE A level standard without allowance for maturation, including those taken at an early stage in the course. If some candidates perform better in modular syllabuses, this may well reflect the positive effects of modularity rather than the standard of assessment. (p4)

However there is not universal belief in the invariant nature of A level standards in general and particularly where modular schemes are concerned. There is much in current (non-modular) A level syllabuses which would not be recognised by an A level candidate of 30 years ago. This does not necessarily reflect a variation in standards, merely the changing nature of the world about us - 30 years ago for example computers were in their infancy! No more should a change in assessment necessarily bring about any alteration of standards.

That there is still much concern is evidenced by the 1995 Interim Dearing Report on the Review of 16-19 Qualifications. This comments in reply to a question concerning the demand of modular syllabuses:

Whilst SCAA and ACAC have an extensive monitoring programme I propose that in consultation with SCAA and ACAC and the GCE Examining Boards, the Review should commission further research to examine the facts. (p17)

There is little doubt that modular schemes have both defenders and detractors within the teaching profession, and it is the purpose of this thesis to investigate those claims of both sides in the argument in order to establish what has been gained, or indeed lost, as the result of the implementation of modular schemes. What exactly is meant by a "change in standards", for it is too simplistic to claim that candidates are gaining different grades from those expected. That merely begs the question "why?".

The Relevance of Examinations

The uses to which evaluative assessment or examinations have been put have changed considerably since their inception in the middle of the nineteenth century. Although

initially matriculation was the first step in degree qualification, examinations were soon seen as a method of selecting candidates for progression to the professions (prominent amongst them the civil or armed services) and there grew an inter-dependence between school curricula and examination syllabuses which has never been broken. A number of other, more wide-ranging functions for examinations have been adduced. Brereton, writing in 1944, cites two - "as a means of stimulating students to effort" and "as a means of testing and classifying students". Within the former of these purposes he stresses the positive educational effect of concentrating both teaching and student effort along pre-defined channels. Dewey (1930), over a decade earlier, had stressed the need for the linking of activities (educational or otherwise) so that examinations became not a termination but an integral part of a continuing process. However, the competitive element was very strong as the School and Higher School Certificate were used as the selection instruments for a number of institutions including those of higher education and the armed services. In general they were a means to an end, not an end in themselves.

When the GCSE was implemented in 1988, it was stressed that its purpose was to enable candidates to show what 'they know, understand and can do'; and the examinations were designed to give candidates as much opportunity as possible to demonstrate achievement. Increasingly the 16+ assessment has been seen less of a school leaving examination than as a means of progression, vide Dewey, to the next stage of education, even though this may take place within the 'world of work'. Many of the assessment techniques pioneered by the CSE examinations were adopted for GCSE use and were eventually accepted as a valid. The changes in these examinations were bound to reflect on the 18+ provision.

The Higginson report 'Advancing A Levels' (1988) quoting from a 1985 white paper commends six examination objectives:

to raise standards; to support improvements in the curriculum and its delivery; to provide clear aims for teachers and pupils to the benefit of higher education and employers; to record proven achievement; to promote the measurement of achievement based on what candidates know, understand and can do; to broaden the studies of pupils in the 4th and 5th secondary years and of 6th form students. (p6)

The first of these is insubstantive, the second suggests that examinations are to be instruments of innovation (though, in fact, they are often the initiators), the third hints at

motivation, the fourth selection. The fifth, the promotion of different assessment methods, again hints at innovation, and lastly examinations are to broaden the curriculum - the very opposite of what A level examinations are often accused (e.g. Kingdon, 1991; Schools Council Working Paper 45, 1972).

If all these objectives are to be fulfilled, then there can be little doubt of the symbiosis between examinations and the curriculum. But it does beg the question - in which of these objectives should comparability lie? Should all examinations equally seek to raise standards, innovate and broaden the curriculum? In practice, comparability means comparability of evaluation, that the measurement of attainment of a number of candidates in a given subject at a given level under different assessment regimes should produce the same outcome - in short, equivalence of grading standards.

Additionally there is the purpose of control. Brereton suggests examinations aid control of pupil by teacher, but the nature of control has changed over the last decade. There is still the controlling effect of motivation, but to this has been added the control of the educational system as a whole (Broadfoot, 1996). By dictating a large part of the content of the A level examination syllabus, including the assessment scheme, by effectively licensing individual examinations and pre-defining the lines along which all developments should proceed, the (now) statutory regulatory authority exerts a restriction over the post-compulsory curriculum which would have been impossible pre-1988. The publication of various league tables of examination results categorising school performance on the (crude) basis of the quality of GCSE and GCE grades has also had a powerful effect on schools' attitude to examinations. Examination results are used not only to evaluate candidates' performances, but also institutional performance, and increasingly that of teachers within the institutions.

Summed up by Broadfoot (ibid.) under the three themes of competence, competition and control, she refers to the "pervasive and characteristic role of formal evaluative techniques in contemporary education systems" and it is undoubtedly important to consider the social and cultural background under which the various assessment schemes have evolved and why the changes in A level assessment have taken the form they have. Deciding which social purpose, if any, should determine comparability is problematic. Issues of validity become pre-eminent - if one examination is a better predictor of degree results than another, does this make the examinations

incomparable? Using such a yardstick it may be easy to compare subjects at A level, but the question would be transferred to comparability of degrees. At best this is unhelpful.

The key issue is the fifth of the educational objectives listed by Higginson, namely to promote the measurement of achievement based on what candidates can actually do. Comparability thus centres on asking candidates to perform equivalent tasks for which the reward is demonstrably the same for equal performances. Not only must demand be uniform for comparable examinations so must the evaluation metric. The role of examiners in this is fundamental. They are responsible for setting question papers and marks schemes which, year on year and within a subject are required to be of equal demand and they are also required to set grade boundaries which reward equal levels of attainment year on year and within subject by giving the same grade.

Since 1951, when subject based GCEs were introduced, the outcome of the "test" (whether a percentage or grade) is a measure of attainment within that subject area which is delimited by the published examination syllabus. Thus no two examinations can ever be exactly equivalent because syllabuses always differ. The differences may lie in content, emphasis, philosophy or assessment, even within the same subject area. Strict equivalence has temporal as well as syllabus bounds, and then only within the limits of marking reliability. Orthodox examination regimes, within option choice, do at least guarantee equivalence for a given cohort. Within a modular scheme it is perfectly possible that two candidates, certificated at the same time and for the same syllabus, have never sat the same examination.

If each examination existed as an entirely independent entity, the question of comparability would not be meaningful and normative grading could be used. That it has significance is a natural consequence of the employment of examination results. Although the causal relationship between examinations and the resultant usage cannot be denied, to employ this relationship in determining comparability would be wrong. If, for example, their primary use is seen as an entrance qualification to some other type of education or training then the predictive quality would be of utmost importance. Thus it would be tempting to compare examinations on the basis of their effectiveness as predictors of success in post-A level education (already notoriously bad regarding degree results: q.v. Christie and Forrest, 1981), although the time delay involved would render such a comparison meaningless. This would be to disregard the more important preparatory aspect of the various syllabuses. Already manifest is one of the more

invidious aspects of published league tables; namely the switching of centres from one Board/examination to another on the basis of so-called "easier" examinations in order to improve league table positions. The evidence for this, though often anecdotal, is seen in the variation in the numbers of candidates entering for each Group's/Board's examinations (GCSE Inter-Group Statistics and GCE Inter-Board Statistics, all years from 1990 onwards) and the changes which have occurred since the advent of performance tables. This is hard to justify in fact, harder still in educational terms. Sometimes the vindication for such a decision is that "a higher number of As" is awarded by one Board over another. Comparability is thus determined by the raw grade distributions un-moderated by pattern of entry. Alternatively the reason may be given that a particular cohort obtained higher grades in one subject over another - a subject pairs comparison. Ignored are the other contributory factors to examination success - teaching effectiveness (albeit allied to syllabus choice) being probably the most influential within individual centres.

Johnson and Cohen (1983) aver, when distinguishing between two definitions based on achievement and aptitude, that the appropriate definition of grade comparability depends

....on which of a number of possible functions the examination aims to serve. (p1)

Because the objectives of examinations as ideally stated are so diverse, it is the purpose of this research to determine comparability independent of any function except that of evaluation - the end is a grade, not a prognosis or league position, and the means to that end is the A level examination requiring a given programme of study and assessment by pre-determined instruments.

Two Approaches to Comparability

It is important to distinguish between the two types of comparability - statistical or judgemental (Nuttall, Backhouse and Willmott, 1974; Bardell, Forrest and Shoesmith, 1978; Good and Cresswell, 1988) which are rooted as much in the methodologies used to determine them as they are in their definitions. Defined simplistically, statistical comparability is comparability of grades or grade distributions once all variabilities in population have been taken into account. Judgemental comparability is comparability of attainment at given grades as determined by suitably qualified judges, once all

variabilities in the examination have been accounted for. Studies in both these strands should come to the same conclusion for the same syllabuses, but it is never possible adequately to take into account all the variabilities in population or all the variabilities in the examinations themselves and such studies using both strands may have different findings from each strand (see Quinlan, 1993 or Ratcliffe, 1993).

Statistical comparability between any pair of syllabuses generally focuses on the average attainment of a specified group, or groups, of individuals by using some second comparative measure such as another subject score, score on a monitor test or a measure of prior achievement, but these raise issues such as bias and relevance which may not be the same for each syllabus. Such methods also fail to address directly the fundamental difficulty of relating examination demand and performance to grading standards, hence the use of judges. There may be imprecision and some unreliability which relate to grade boundary setting within an examination component and when allied to structural and syllabus differences between examinations invests the task of comparison with a complexity not easy to resolve. As Forrest and Shoesmith (1985) remark with reference to setting grade standards:-

...that standard is nowhere explicitly laid down: it lies instead in the experience and minds of those teachers and administrators whose task it is to carry the standard across time. (p23)

and therein lies the central tenet of this thesis, that awarders are able to perform this task in a consistent manner across time and syllabuses and in a manner which reconciles all those unquantifiable variabilities so that all grades awarded at a given level are accepted as equivalent.

There is no obvious paradigm for examination equivalence - two sticks can be compared in length without requiring the defined length of either, and one is said to be "longer" than the other - the lack of a common dimension makes calibration unfeasible even within the limits of marking reliability. The very terms leniency and severity which are usually applied to the perceived relative difficulties of examinations imply a value judgement which can rarely be justified. The problem is essentially one of measurement. Grades are measured on an ordinal scale, but are often treated as continuous. The closest attempt at calibration was the 1963 Secondary School Examinations Council guidelines which suggested cumulative percentages for each grade; 10% at A, 25% B,

35% C and 50% D with a total pass rate of 70%, but it was emphasised that these were no more than rough indicators. As discussed in the SSEC report which recommended the guidelines (SSEC, 1960), although Boards try to ensure equality of demands of syllabus and examination

they do not normally control the kind of students who will offer themselves for examination (p27)

i.e. the aptitude or ability of individual candidates.

In this thesis attainment and achievement can be regarded as synonymous and defined as the outcome of a specific A level examination as manifested by the grade awarded; aptitude has yet to be defined. In Nuttall, Backhouse and Willmott (1974) the covert definition is:

... the overall ability of a candidate to achieve good grades in the examination (p5)

which is a specific application of that given by the Oxford English dictionary of aptitude as fitness; natural propensity; ability. Psychologists however make a temporal distinction between aptitude and ability - aptitude is "the potential to perform", ability is an enabling faculty now. It could be argued that at the beginning of a course of study a student possesses aptitude, at the end ability and after some relevant test can be said to have demonstrated a given level of attainment. This will reflect ability rather than aptitude. The difference between aptitude and ability is therefore a measure of the effectiveness of a given course of study. Hall (1977) considers general ability as manifest in three forms; general intelligence, general educational ability and general scholastic ability. In a sense, even though this may be too simplistic (q.v. Christie and Forrest op. cit.) the mix is unimportant since without explicit measurement, impossible within the context of this research, aptitude can only be given by proxy, as for example average GCSE grades, with ability in part demonstrated by A level results in cognate subjects. The thesis that attainment is functionally dependent upon aptitude assists in the formation of an equation of outcome (such as a regression equation based on prior achievement), but the interdependence of terms in such an equation should not be ignored; i.e. aptitude may well act as a parameter rather than a separate variable.

The GCSE General Criteria define comparability as:

....the extent to which the same grades in different examinations represent the same or equivalent levels of performance. (p21)

This is clearly an "attainment" definition, and is the umbrella under which equivalence should be determined. However, its self-evidence as a definition gives little insight into what comparability means in practice.

Christie and Forrest (ibid.) are equivocal about comparability definitions based upon aptitude. However, in this thesis, the requirement to compare modular with non-modular examinations in the same subject may invest a (construct) validity to such an approach which could be missing with between subject comparisons.

It would be easy to guarantee equivalence in one sense, by norm-relating the grade distributions of all syllabuses. Indeed there is a powerful argument for doing just that, especially for the large entry subjects at GCSE. The definition of comparability taken from Nuttall, Backhouse and Willmott:-

...we can see no logical reason why, if a large group of candidates representative of the population took, for example, both English and mathematics, their average grades should not be the same. (p12)

was applied to O level and CSE examinations. A sufficient (though not necessary) condition for this to hold would be to norm relate the grade distributions for large entry subjects and would endorse the argument above. However, at A level, although the definition could hold, it is difficult to see that the underlying stated assumption would; the non-homogeneous, self-selecting and often small, candidature is unlikely to be "representative of the population". For modular and non-modular schemes (at least within an individual board) candidature is mutually exclusive by regulation; there may also be some fundamental difference since the two populations have chosen to follow separate syllabuses. An interesting aside relating to the two subjects chosen as examples is the necessity of ensuring strict representativeness. A recent study (Stobart et al 1992) would indicate that the dominance of the group by one gender would materially affect the group average. If the group were predominately male, then because boys do better than girls at mathematics (though the gap is closing) and worse at English (and the gap in English performance between males and females is widening)

then the relationship between the mean performances of the group for the two subjects would be very different from that obtained should females dominate. There may also be an issue relating to the quality of teaching of the two subjects, especially given the relative difficulty in recruiting suitably qualified teachers in mathematics.

The Problem

Modular (or "structured") schemes of assessment involve three processes which differ from the conventional approach:-

(a) that module examinations may not necessarily be terminal
(b) they always involve some choice, including the choice to resit;
(c) that module results are strictly additive under the UMS[1] transformation.

Different modular schemes currently vary in their approach to all of these; the application of (a) has led to the resitting of some modules for some examinations, (b) the suggestion that some routes are easier than others and of (c) to several different methods of combination. The Code of Practice for GCE A and AS Examinations restricts certain choices; terminal examinations must contribute at least 30% of the marks and each module should be worth at least 15% of the total assessment. It is the purpose of this research to investigate definitions and methods of determining comparability and apply them to ascertain whether modular examinations can be considered equivalent to their conventional counter-parts.

The reductionist approach of modular schemes to grading, though becoming more popular across all subjects (and is embodied in the Code of Practice) brings with it fundamental changes. Rarely will all of a candidate's work be available for inspection so holistic judgements become unviable and the longitudinal nature of the examination imposes a commitment to standardisation not so necessary for conventional examining. It perhaps strengthens arguments for criterion referencing albeit difficult for some subjects since statistical data are less reliable indicators of performance (because of the different populations taking modules) than they are for conventional schemes and attainment of common criteria may help to anchor standards from session to session.

[1] UMS, or uniform mark score, is the name given to the standardised module score which involves a transformation of module raw marks on to a common scale for aggregation purposes.

Grade descriptors (where they exist) would be crude examples of such criteria.

Attitudes towards marking may also have to change with a greater willingness of the more subjective examiners to utilise the whole of the mark range. The attenuation of the mark bandwidths for, say, English could have profound effects on outcome when compared with, say, mathematics examinations were a universal uniform marks scheme to be adopted. Grading for modular schemes can exhibit characteristics of both continuum and state models in domain-referenced measurement. Although the achievement domain may vary for the "whole subject" grade, within individual modules it does not and so grading can also be said to be limen-referenced, a rather ad-hoc mixture of criterion and norm-referenced grading (see Christie and Forrest op. cit.). The difficulty in pinning down precisely the grading procedures employed relative to conventional measurement models adds to the complexity of defining equivalence.

In a sense it is not necessary precisely to define any variations in grading procedure. If we can make the assumption that grading standards are the same, irrespective of grading scheme, then the task becomes one of looking at the differences between schemes which may, or may not, impinge upon the grading process and discovering if there are any inequalities in outcome which are unexplainable except by differential standards. This applies not only within a modular syllabus but also between a modular syllabus and a linear syllabus in the same subject.

The question to be addressed then becomes not one of comparability, but one of how valid each of the schemes is as a method of assessment. Key to this is a consideration of the threats to validity usually defined as construct underrepresentation and construct-irrelevant variance (Brualdi, 1999). Essentially what needs to be considered is whether the tasks which are evaluated in the two schemes are the same and whether the nature of the questions set or the assessment structure itself inherently lead to different expectations of outcomes i.e. grades. In order to consider the latter, the variabilities between modular and linear schemes need to be exposed and their effect quantified. In this way, should differences be found in attainment by equivalent candidates as defined by, say, a prior achievement measure, it may be possible to determine whether such differences render either scheme invalid.

Variabilities are of two types; illegitimate variability is one which must be considered by awarders when setting their grade boundaries and legitimate variability which might go

far in explaining differences in candidate performance but which plays no part in the judgemental process per se, although they may go some way to explaining a better than expected performance. An example of illegitimate variability might be question demand where judges may have to reconcile high attainment on easy material with lower attainment on very much more difficult content. Legitimate variability may, for example, be due to pupil motivation or whether the candidate is resitting, but this would not be a consideration in the grading process where the only concern is determining a grade on the evidence of achievement provided by scripts and coursework.

Whilst it is not possible within this thesis directly to consider legitimate effects due to, say, motivation (of both pupils and teachers), the more mechanical elements of the examining process can be addressed. By assessing the effects of re-sits, of examination session, of the effect of taking further modules and indeed gender, it is possible to estimate the source of differential performances and possibly quantify them. If the quantification of the effect of legitimate variabilities as defined here can explain performance differential then grading standards may be considered sound. The discussion would then move to a debate on what constitutes a legitimate variability.

Methodology

It is proposed that the research centres on one subject area, namely Mathematics, primarily because of the few established modular schemes, one of the most popular is the OCEAC MEI Structured Mathematics Scheme which not only provides a stable data base, but one with an increasing number of candidates. Although other subject areas may help to illuminate certain of the inter-module relationships, inevitably clouded by the resit factor in a well established scheme, most of the work focuses on those candidates who were entered for certification in the summer of 1994.

This thesis takes as its starting point one major assumption - that of comparability of grading standards between linear and modular schemes of assessment. By investigating in some depth the variabilities referred to above, both within the modular scheme and between the modular and a chosen linear syllabus in the same subject area, an attempt is made to find and quantify sources of difference and compare them with expectations. It is then possible to learn whether the initial assumption has been violated.

The variabilities can be categorised under two headings, those occurring either within a syllabus or between syllabuses in the same subject area. Because the flexibility of modular schemes introduces an element of uncertainty regarding the equivalence of modules not seen within conventional schemes, the first part of this investigation centres on an analysis of the modular data to determine and quantify *within syllabus* legitimate variability whilst the second concentrates on an analysis of syllabuses, question papers and mark schemes as well as item level data for both conventional and modular examinations in an attempt to find any inconsistencies which might indicate a failure of the underlying assumption. Then, by using a measure of prior achievement and looking at differences in attainment between equivalent candidates as defined by the measure, an attempt is made to reconcile such differences with those legitimate variabilities previously found. In this way it is possible to assume that all sources of illegitimate variabilities, including the difficult area of question structure and demand, are subsumed by the grading process - unless proved otherwise.

The statistical methods used for the analyses are, for the most part, tried and tested in the furthering of examination comparability. However, this has been at subject level, and the application of the methodologies to modules has provided an insight into the working of a modular examination which was previously a source of speculation. Such applications can be extended to other modular schemes, and inevitably so too can some of the conclusions, although a number of the findings are subject specific.

Firstly, the research concentrates on the typical behaviour of a modular candidate as exemplified by the chosen scheme. This includes sessional variations, resit patterns and comparability between modules. Multi-level modelling techniques enable us to look beyond the obvious and consider those factors which may impinge upon performance and their variation with centre.

Question level analysis from one modular and one linear scheme for the same year of 1994 enable a direct comparison of the demands within each syllabus and question performance on each of the papers. Insights derived from this analysis enable some tentative conclusions to be drawn about comparative demand.

The final analytical section is concerned with comparison of performance on linear and modular schemes both with each other and across time. Interest focuses on the variance of the scores of candidates at centre level, and across the years, and the tool

for the analysis is again multi-level modelling. Some interesting conclusions can also be drawn from this study.

The two chapters involving multi-level modelling techniques are as much an exploration of those techniques as of the modular syllabus under investigation. For that reason, even where not particularly successful, the methodology is reported for its relevance and potential use in other comparability studies.

Outline for Chapters

Whilst the literature on examination comparability is sparse, with nearly all having been produced either by the boards themselves or SCAA (and its predecessors), there is enough to provide the background necessary for this investigation. Methods of determining examination comparability have been developed and refined over the years, although most are not applicable to the research carried out here since they rely on the input from experienced awarding personnel. Modular schemes have developed in a far from structured manner, and it was not until 1992 that the issues were addressed properly in a report by David Thomson and an aggregation system proposed. This report can be considered seminal in terms of its influence on the subsequent development of modular schemes and the method in which the parts become the whole. Chapter 2 investigates the literature on comparability.

There is a certain amount of disagreement between the A level boards as to which can claim to have set the first modular examinations. However there is little dispute over the existence of modular schemes at GCSE long before they were even considered a suitable regime at A level. This does, however, beg the question "what is a modular examination?". There is no doubt that the modular GCSE courses were very different in their assessment methods from those that are now applied to modular A levels, and that even now "modular" tends to be a term applied to any A level, parts of which can be taken early. By looking at the history of examinations in general and modular schemes in particular, specifically in the arena of public examinations, it is possible not only to contextualise the current regimes, but also to trace their development from their beginnings as a method of taking and assessing degree performance in the 1960s to the applications of today. This is the subject of chapter 3.

Research into examinations is often one of two types. Either the investigator sets up experiments in order to discover or determine particular behaviour patterns (in their widest context), or (s)he sets down some hypothesis which the research is conducted to test out. In the context of modular schemes, although the candidate may find it a journey into the unknown, there can be no question of setting up an experiment to test hypotheses. Pilot studies do work in very specific contexts, but it is unlikely that the whole variety of a two year course of study could be the subject of a single experiment, especially when one considers the importance of the result to the candidate. This research therefore must fall into the latter category, and the fourth chapter concerns itself with the consideration the sources of invalidity and their relationship to a number of factors, or variabilities, relevant to various examination regimes. A hypothesis, which, if not directly testable, at least questions the conventional wisdom that comparability is simply a matter of ascertaining whether a candidate would have obtained the same result had (s)he followed a different syllabus in the same subject is postulated. It is suggested that the dichotomy between the two major assessment regimes under consideration is sufficiently profound to lead to expectations of different outcomes even when the candidates' ability (using some prior measure) is the same, but that these differences are the result of entirely legitimate definable variabilities, and not necessarily threats to validity. Such variabilities would be entirely relevant to the construct and provided that it can be shown that the evaluated elements within each scheme are essentially the same then each scheme must be valid under its own defined construct. The issue then becomes one of belief in the social equivalence of the results.

For example, if it can be shown that candidates tend to get higher grades from a modular scheme, does it necessarily mean that standards have dropped. It may well be that the assumption of equivalent grading standards still holds, but that the assessment regime is more 'enabling' and by setting targets throughout the course and allowing a certain amount of re-taking within the course, candidates are learning more. It could be argued that weaker candidates, through their re-sits, are more proficient at the topics covered earlier in the course than they would be had the examination been taken terminally.

This chapter also outlines in detail the types of scheme currently available and their different aggregation methods and considers the origins of the many differences between conventional and modular schemes.

The thesis centres its attention on one modular syllabus, the MEI Structured Mathematics scheme, not least because it is incontrovertibly modular. An outline of the scheme, its aggregation method and comparable traditional syllabus (used as a control) are given as are details of the awarding procedures which lead to the various grade boundaries being determined. The two strategies of awarding grades, though dictated by surface factors do disguise quite significant differences in application. On the one hand, traditional schemes allow for "regression" in the aggregation, that is the tendency of candidates to perform less well the more components are taken into account. There are practical problems as well as philosophical objections to attempting to make this allowance in modular schemes (though this was not always the case and rather odd estimates emerged for use as regression allowances) and each module must be graded on a "stand-alone" basis. These differences can cause difficulties where syllabuses declare themselves as both modular and linear. This thesis does not accept the proposition that a syllabus can be both: even if all modules are taken terminally, the examinations themselves and aggregation processes are still sufficient to underpin the basic dichotomy, as there are if certain components in a traditional scheme are taken early in the course.

Investigation of the sources of legitimate variability is the focus of Chapters 5 and 6 and of the comparability of modules. Differences in module standards can arise in two ways; either as a displacement, all the candidates obtain consistently better (or worse) marks on two different modules, in which case the correlation should remain fairly high; or standards are applied inconsistently across boundaries in which case the correlation would probably be somewhat lower given the method of aggregation on a uniform mark scale. However, compounding these two fundamental sources of difference are the different occasions on which the same module may be taken, within different populations each time, and the choice of modules.

Chapter 5 is concerned with the basic data of the chosen modular scheme. Details of resit patterns, module choice and performance are all given as well as consideration of the correlations between modules. Whilst it is unlikely that similar patterns of behaviour would directly translate into other modular subjects because of the uniqueness of the further variant, some of the richness of the variety should (where allowed) and also the effects of the different choices, observable here in resit patterns and seasonal variations.

It is this effect with which chapter 6 concerns itself. The approach is one of multi-level analysis of the data set aimed at determining the equivalence of modules, the effect of gender and re-sits. To this end a more restricted data set is used, namely that containing only the modules which counted towards certification for the single subject mathematics (i.e. six modules per candidate) rather than the overview of all modules taken by candidates who were certificated (i.e. including extra modules which may have been used for a Further A or AS qualification). Practical difficulties prevent all modules being analysed together, but it is possible not only to look at various combinations of modules, but also to investigate the effects of the different factors at individual module level. Also, with this restricted dataset it is possible to investigate internal consistency, which is a measure of reliability of the examination, and the achieved weight of each module.

The investigation presented in Chapter 7 focuses on the syllabuses, their differences and similarities and the way the two schemes of assessment, modular and linear, each perform in terms of sampling the syllabuses in the question papers. The thrust of the analysis is to determine whether there is construct under-representation in either the modular or linear schemes. This investigation requires a value judgement in order that meaningful comparisons can be made. The function and value of a mark can only be determined within the context in which it is awarded. For example, in a four part mathematics question which has attached to it 10 marks, the first, second and third parts may attract 3 marks each and the last just one. However the function of the first 3 marks is to enable the awarding of the second 3 (i.e. without the first part the second cannot be completed) and so on. This is rather different from, say, a 10 mark question which consists of two, unstructured independent parts of 5 marks each for which the most likely scores are 0, 5 or 10. In structured questions designed specifically so that the first part is accessible to all, the last to very few, it is often easier for weaker candidates to score at least some marks than with the more inaccessible unstructured form. In the modular scheme considered, the aim is for all questions to be structured and this makes comparison with traditional papers even more difficult. Using the mark schemes as a guide it is, however, possible to put a value on each question. Given this data a comparison of the relative requirements of the various question papers can be made.

Allied to the evaluation of question papers and mark schemes is an item analysis of a random sample of scripts from both modular and non-modular schemes. This contextualises the mark schemes and allows a more direct comparison of abilities of candidates. Within each part of the examination (either component or module) a picture

of performances in the areas of skills and knowledge can be constructed. In this way, a picture of the effectiveness of the two schemes in allowing candidates to demonstrate their various abilities can be constructed.

As has been argued earlier in this chapter, the aim of this research is not just to suggest that it might be easier to gain high grades under one type of assessment scheme. The truth, or otherwise, of the assertion of apparent lowering of standards has to be proved. Since one of the best indicators of performance at 18+ is performance at 16+, a matching of GCSE data and A level data for the same candidates in the subjects under investigation is ideally required. Subjected to the appropriate analyses it is possible to gauge the differences in GCE performance based on expectations derived from GCSE results. In chapter 8 the results of such a multi-level analysis are presented, using results from two consecutive years. So there is not only a comparison between a linear and modular scheme, there is the added dimension of a year-on-year analysis.

The final chapter aims to summarise the major discoveries and present such conclusions as can be made on the effect of the variabilities which have been defined and analysed.. In particular, there is discussion focusing on the robustness of the underlying assumption of equivalence of grading standards. The social and cultural context in which modularity has arisen is considered as is the generalisability of the findings. The major points are summarised in propositions which embody the research. The focus of this chapter is wider than the research itself, considering issues of both assessment and control as well as the comparability with which the investigation concerned itself.

CHAPTER 2

A More Forcible Word - The Literature on Comparability

Wood (1991) observes in his survey of research :-

Modularisation is merely question choice writ large. It exists to accommodate personal taste and to make syllabus coverage manageable. (p16)

If it were that simple it would be possible to look to the literature on question choice to inform the research on the comparability of modular and non-modular schemes. There are quite clearly fundamental differences between the two, the very nature of the assessment and aggregation within and between components ensures this. However as Thomson (1992) says of modular schemes:-

Because it is a new departure little experience exists of what exactly the philosophical, pedagogical and technical implications of "modularisation" might be. (p2)

Two years after the publication of his report on modular aggregation and comparability, experience has been gained, but little has been written of the rationale which now governs the awarding of modular schemes or of the impact of such rationale on comparability.

Thomson (ibid.) has also taken Broadfoot's categorisation of the Psychometric and Educationist models as a paradigm for the traditional and modular approaches to examining. Certainly many features of the Educationist model, with its emphasis on formative aspects of assessment rather than summative measurement, incorporate the ideals towards which compilers of modular examinations strive, and there is evidence that modular schemes are thought to enhance pupil motivation (Gray, 1992; Nickson, 1994), a major element, especially amongst the less able. Whether the essentially different approach of modular examining can ever produce comparable results to those from orthodox regimes depends, at least in part, on what is meant by "comparability".

Pollitt (1993) defines this as "the equivalence of identical labels" and appeals to Campbell's laws of measurement to show that the requirements "specify our task in public examining." Put simply, Pollitt states that these require that:

all Bs are equal in value, A>B>C etc and we can add up the separate results somehow.

Whilst all three requirements may be desirable, any direct attempt to prove that a C in Physics is equivalent to a C in drama for example is probably doomed to failure (Goldstein and Cresswell, 1996). If one accepts the additivity of grades, then much of the recent work on aggregation of components or modules whether on differentiated or undifferentiated examinations (q.v. Good and Cresswell, 1988; Backhouse, 1976; or Thomson, 1992) is rendered meaningless. In fact grades are too crude a measure to be used within an examination, and too different to be used between examinations. The distances between grades have no meaning per se, and therefore grades should not be assumed to constitute measures on an interval scale, though they are usually taken as such. But they are measures on an ordinal scale, and comparability requires that the ordinal scales are effectively the same i.e the laws of reflexivity, symmetry and transitivity on the relation of equivalence and of irreflexivity, asymmetry and transitivity for the relation of "greater than" hold between the grades of one examination and another. UCAS perpetuate the use of grades as points, and indeed many research studies find assigning numerals to grades allows analytical techniques to be employed where, strictly, they are not appropriate.

Definitions of Comparability

In his book on Assessment and Testing (1991), Wood asserts (p117)

A special case of scaling (or equating) is the very British problem of comparability, though it may not seem so at first glance.

This statement makes the assumption that the determination of comparability is not only possible and meaningful, but that if only one could discover the correct calibration instrument, equivalent outcomes would be assured.

However, a number of authors have suggested that the assumption that comparability is a meaningful concept does not always hold (Newton, 1997) and that the methodologies, either qualitatively or quantitatively, to determine its nature

(Goldstein and Cresswell, 1996) are implicit in unsatisfactory definitions as to its meaning.

Whilst the comparability of examinations in different subjects may, in practice, only exist as an abstract concept, there is a need to consider if the same is true of examinations within the same cognitive domain. If equality can exist in some form, then it must be definable and testable

If we take as our starting point the 'social' definition of comparability (Cresswell, 1996, p79):

Two examinations have comparable standards if candidates for one of them receive the same grades as candidates for the other whose assessed attainments are accorded equivalent value by awarders accepted as competent to make such judgements by all interested certificate users.

In fact, 'all interested certificate users' have very little choice. The investment in, specifically, A level examinations by boards, examination centres and most importantly, candidates, means that institutes of higher education (to which the majority of candidates aspire) have little choice but to accept a subject grade at face value, irrespective of board or assessment scheme. The UCAS form, which each prospective university entrant must complete, takes no cognisance of scheme of assessment (although the syllabus board of origin is included), and if a student wishes to advertise early module results it is his/her prerogative to do so, but it is not compulsory.

So we have a situation where, practically, all examinations in a given subject at a given level are deemed comparable, because it is in very few peoples' interest and impractical to do otherwise. In the same way all such examinations are rendered valid because (Cronbach, 1971) the

interpretation of data arising from a specified procedure

i.e. an A level examination, is the same for the same grade, irrespective of origin.

However such arguments are relatively superficial, since there may well be a gap between what can be inferred from the assessment and what has been inferred (see Messick, 1989). If all reasonable inferences are valid for two schemes of assessment within the same subject (at the same level), then they must be comparable. If some inferences are not valid for one scheme or another then there is a lack of comparability between the schemes.

This ties in with the definition of comparability which relies on the expectation of similar outcomes (Cresswell, 1996, p70) i.e.

had a group of examinees followed another board's syllabus and taken its examinations, they might reasonably have expected to produce the same distribution of grades

with the all that this implies - specifically that learning outcomes from the course of study should be the same, whether tested or not.

Cresswell (ibid.) expands on three major issues which arise from this definition:

(i) that the examination itself, the schools and the reaction of the students to these variables will affect performance in unpredictable ways
(ii) the content, or value, of what is assessed must be the same
(iii) that distributions should be the same for identifiable sub-groups of candidates as, for example, boys and girls.

If these three factors are compared with the usual trinity of validity definitions i.e. construct, concurrent (predictive is not an issue in this particular debate, although preparation for a particular course of study may well be) and content and use Ebel's (1965, p380) definitions then (i) may be (loosely) equated with the psychological qualities measured by the test i.e. construct validity; (ii) is equivalent to content validity; and (iii) is certainly concerned with the relation of the test scores to an accepted criterion of performance i.e. grade, and this is the definition of concurrent validity.

Further, if these three factors are re-defined as validity inferences, then a comparability study should aim to examine the extent to which they are true for the syllabuses under consideration. Not all inferences are testable, and this is the main weakness in the determination of comparability, much has to be taken on trust; it is, for example, impossible to measure directly the effects of the syllabus itself (and reference tests are not the answer (see Goldstein and Cresswell, op. cit)). One way to approach consideration of equivalent validity might be to consider the two factors which would together present a threat to that assumption, construct under-representation and construct irrelevant variance.

There is a further complication when one of the schemes of assessment is modular. Since each is a self-contained examination in its own right, there is the assumption that each and every version of the examination is comparable, because, unlike linear schemes where standards are assumed robust with in a given year's cohort (though it must be noted that the internal consistency of examinations where options are available is far from assured), the variablity of modules within each candidate's portfolio can be such that differential validity is possible. Therefore the inferences which may be drawn from a given grade in a given subject must also be seen to be robust from within the syllabus.

However, much of the work on comparability has assumed, conceptually, its existence and a number of papers quote, in some form, the 1977 words of the Expenditure Committee of the House of Commons, (though they probably originate in Nuttall, Backhouse and Willmott, 1974)

....that examining boards strive to ensure that standards be kept as similar as possible:

> *(a) between various boards*
> *(b) between the various subjects*
> *(c) from one year to the next*
> *(d) between the various mode 2 and mode 3 schemes.* (p11)

Given that from 1994, the GCSE mode 2 and mode 3 schemes have disappeared (a hangover from the dual certification of GCE and CSE which, when combined in 1988, still left certain CSE practises in place including those examination curricula which were

school devised and moderated), it is reasonable to paraphrase (d) to encompass the different examination structures within a board/group which exist at both GCE and GCSE to certificate what is ostensibly the same subject. This is one type of comparability with which this thesis is concerned. It is also important to recognise the added importance of year on year comparability in modular schemes. The usual model for this type of examining will offer at least four, and possibly six, opportunities within the usual two year A level course to take and/or resit at least some of the modules. The implications of non-comparability year on year would be sufficient to cast doubt upon the validity of such an examination regime. It is a fundamental premise that individual modules are comparable year on year since the candidature often include results from the same module number taken at different module sessions. (More will be said on this later).

Because it may be inferred from the positive choices made by centres to transfer from traditional examining methods to modular schemes, that there may be some essential difference between the two populations of students, before dismissing the issues of mode comparability, it is worth taking as a caveat the point made by Bloomfield, Dobby and Duckworth (1977) in their inter-mode comparability study:-

Hence the probability that pupils are not assigned randomly to examinations of two different modes implies that the results of this study will have to be interpreted with considerable care. (p17)

with "structures" replacing "modes".

But it must be emphasised that two examinations may be deemed comparable without being equal. Until 1983 when the GCE Examining Boards of England, Wales and Northern Ireland published its "statement of agreed Common cores in certain subjects at the Advanced Level of the General Certificate of Education", it was possible that two subjects with the same syllabus title had no common content. Early workers in the comparability field had first to ensure that there was sufficient material to warrant "within subject" comparison.

In 1979 the Schools Council published the first of two occasional papers which were the product of the Council's Forum on Comparability. This focuses on the complexities

which arise because examinations can never be altogether comparable; and the reasons why this is so need to be explained to a public who subscribe to the generally held view that the grades awarded by the different examining boards are, in some undefined way, the same. Here, in occasional paper 1, is found a statement of the underlying assumption of most, if not all, comparability studies:-

A necessary condition for the achievement of explicit comparability between different examinations in terms of skills reflected by equivalent grades, is that the grades awarded to individual candidates who sat the same examination with the same result should reflect similar performances. (p13)

Clearly if two candidates awarded the same grade from the same examination cannot be demonstrated as being comparable in any meaningful way (except presumably marks total which of itself means nothing until linked, or referenced, to a question paper and mark scheme) the problems which would arise in the task of determining between exam comparability become insurmountable. If, for example, a candidate gaining 50% of the marks total was awarded a grade C, then it is theoretically possible for two grade C candidates to have no common marks, within the same examination. This does not happen, or at least not in mathematics which has highly correlated components. Whilst it may be contentious to suggest that every grade C candidate can succeed at the same tasks, it is probably true that most grade C candidates will fail on some common tasks, especially in a hierarchical subject such as mathematics. This is evidenced by the use of 'benchmark questions' in awarding meetings which are used as discriminators between grades. In other words it is, in part, what they cannot do which defines them as well as what they can do. It is, however, important to demonstrate that at any grade, candidates have shown sufficient common knowledge and skills for that grade to be indicative of a level of attainment, as it is even more relevant when comparing standards across examinations. 'Grade C-ness' should be definable and meaningful before such comparisons can be made. Without such a concept, however variable, it is difficult to see how awarders could ever make a judgement as to where a given grade boundary should lie. That such judgements are made and deemed comparable gives rise to the notion of social comparability.

Strictly the stated requirement relates only to reliability, but this is so closely allied with validity that they are usually considered together as fundamental requisites for an effective examination. Ebel (1965) puts it succinctly:-

validity...is the accuracy with which a test measures what it is intended to measure. This is in contrast with reliability which can be defined as the accuracy with which the test measures whatever it does measure. (p389)

He also goes on to say:-

Reliability[1] *is a necessary condition for validity, but it is not a sufficient condition.* (p389)

Indeed, Deale (1977), who says much the same thing emphasises this by stating that "validity is the first consideration" because "to be valid, a test *must* be reliable *and also* satisfy other requirements." However, classic psychometric theory is concerned primarily with the establishment of a "true score", which is usually defined as the actual score on a given test plus an error term, not a validity term. In fact validity in most examination research is assumed.

The issue of the indivisibility of validity and reliability is not clear cut - essay questions, for example, are difficult to mark reliably because much of the judgement about the quality of the work is subjective, and the yardstick for those judgements is founded in the experience and expertise of the examiner. Reliability can be so low as to render the examination results invalid - though the test itself may have been perfectly valid.

Messick's classic work on validity (in Educational Measurement, 1989) is built on a unitary theory i.e. the three conceptions of content, predictive and concurrent criterion-related and construct (Cronbach and Meehl, 1955) are compressed within the 'social settings in which they are made' (Wiliam, 1993). Nuttall (1987) defines validity as:

an inductive summary of both the existing evidence for and the potential consequences of test interpretation and use. (p110)

---

1 Since measures of reliability are usually continuous between 0 and 1, this presumably implies a positive, non-zero reliability.

In terms of this thesis, if it could be proved that modular schemes are valid under this definition, then their 'generalisability' to all possible uses of the results i.e. the universality of application, would be assured. In practice, it would be enough to show that modularity is at least as useful as linearity for the purposes to which these assessments are usually put. The link between this and the aforementioned social comparability definition is unavoidable since within this is the implicit assumption of equivalent validity.

Ebel (1965) lists ten different types of validity, although he notes that they are "not all distinctly different". However this type of taxonomy is out of favour and Wood (1991) points out that validity studies are rare in British examining boards, although estimates of internal reliability coefficients in the form of Cronbach's alpha or Backhouse's P are routinely made.

There is however another aspect of reliability which needs to be highlighted, and it is found in a definition by Ebel (1965):

The reliability coefficient for a set of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores on an equivalent test obtained independently from the members of the same group. (p311)

The problem with this is that there is never an entirely equivalent test and in the realm of public examinations there are a number of reasons why the same set of examinees might not produce the same outcome from different, though similar, examinations - they would have been prepared for one examination only, practice papers would have been different, assessment may be by different methods, weighting of components may be different and so on.

The School's Council paper on comparability (1979) also defines comparability between subjects with the idea that:

it is usually assumed that a group of students of a given level of general ability should be expected to achieve the same average grade in each subject they all attempt (p10)

This assumption of common ability leading to common outcomes underpins much of the comparability work which relies on comparisons of two variants of an examination with

common others e.g. subject pairs analysis, though it probably poses more questions than it answers, a fact which is recognised by the authors. It is not an assertion which many would defend with any rigour. Correlations between subjects are often low and ability demonstrated in one subject is an imperfect indicator of aptitude in another. Just because a candidate is good at mathematics, say, is no reason why they should demonstrate the same level of ability in, say, art & design (two subjects whose correlations are notoriously low).

They do however give a more general definition of within subject comparability - and though posited for between board comparability can be applied equally well to different syllabuses offered by the same board, because the essential point is the diversity of syllabuses within a given subject - namely:-

the expectation is that had a group of examinees followed another board's syllabus and taken its examination, they might reasonably be expected to have obtained the same average grade. (p11)

This is a generalisation of the 'reliability' definition, and has the merit of being simple and relatively uncontroversial, given a random sample of sufficient size, and is certainly the goal for which examination boards aim. It also avoids the pitfalls associated with the use of "ability", itself a difficult concept to define, especially in examination terms. It may be enough to accept the two manifestations which are measurable (at least by examination boards) that of general ability, as for example given by an average GCSE score, and subject specific ability, which might be exemplified by an A level result in that subject. This would be in line with Gardner's (1983) 'multiple intelligences' though, as defined, they are more general than the two defined here.

But there are a number of different ways in which a mean grade could remain invariant and comparability could not be assumed. A more rigorous definition might consider the idea of congruence which would imply that the rank order of candidates would remain invariant under the two examination regimes as would the grade distribution. These two conditions would separately be necessary for comparability, together would also be sufficient. Averages would be the same, but just as importantly so would the candidates' grade, though not necessarily the marks. Any study in comparability should attempt, at least by proxy, to determine whether these two conditions hold. Since the same cohort

could not take two examinations the difficulty with this type of definition lies with determining an equivalent set of candidates.

Two further variations of definition are supplied by Orr and Nuttall (1983) in the second occasional paper:-

There appear to be two basic notions of comparability: one is that different examinations should test and reward in a similar manner the attainment of the same specified skills; the other is that for a given level of ability, the expectation of a candidate achieving a specific grade should be constant (across syllabuses or across time). The first notion may be seen as a narrow interpretation of comparability and the second as a broad interpretation. (p20)

This is clearly aimed at within subject comparability, although some skills cut across subject boundaries even though not explicitly tested. These notions are analogous to the Cresswell (1996) definitions of 'equal-attainment' comparability and 'value-added' comparability, both of which are subsumed by the social definition of comparability.

However, the two propositions can be characterised in ways other than "narrow" and "broad". As is explained (ibid.) there are two possible and distinct methods of awarding grades:

> (i) criteria-referenced grading where, in order to achieve a given grade candidates have to show "pre-determined levels of competence in specified aspects of the subject concerned." This defines an absolute measure for standard setting. Christie and Forrest (1981) prefer to use the term domain-referenced grading, where criteria exist in a domain of behaviours for which a status can be defined. This follows the ideas of Glaser (1963), the first proponent of criterion-referenced tests, who thought in terms of a "continuum of knowledge acquisition ranging from no proficiency at all to perfect performance". It is within this continuum that criterion levels would be set such that mastery of the levels would provide a measure of student performance. It would seem that this may have been the original model for the National Curriculum as first proposed. However, according to Glass (1978), criteria were then used to set "cut-off scores", rather as used in the context of signal detection. The difference

between the two is probably more semantic than real, hence the term 'criteria-related', and in the context of public examinations the criteria are specified by the mark scheme through which marks are allocated. In the sense that a candidate is either awarded the marks or not, the criteria are defined in the signal-detection sense:

(ii) norm-referenced grading, which is purely relative, is not defined in the sense of a single mean or average, but determines grade boundaries based on given percentages of candidates in each grade. In 1963, when Advanced level was introduced, a set of percentages was suggested as a guideline - 10% at A rising to the 70% who would be expected to gain a pass grade. In general, such grading is usually related to previous grade distributions in the same syllabus and has also been defined as 'cohort-referenced' (Wiliam, 1996)

As Orr and Nuttall (1983) point out:

In practice, grading of public examinations relies on both norm-based and criteria-based considerations. (p13)

This complexity poses problems for those studying comparability. French, Slater, Vassiloglou and Willmott (1987) would go further:

....it seems to us that public examinations can be neither norm- nor criterion-referenced in the strict sense. (p15)

Christie and Forrest in their 1981 book are the first to apply the term "limen referenced assessment" to the process of grading public examinations which they admirably dub a "fuzzy process" based on the subjective judgement of awarders. Thomson (1992) avers, following the ideas of French et al. (op. cit.), that it is therefore a form of criterion referencing where the criterion is a standard which exists in the mind of an experienced examiner. In as much as it is not specifically norm-referenced this is true, but such a definition is of no constructive help in determining comparability, except possibly in the context of a cross-moderation study. The inability of so many analysts to define precisely the form of assessment which characterises public examinations in Britain stems from the two distinct measurement processes which combine for the award of a grade to the individual:-

(i) marking (analogous with counting) and aggregation, which is precise within the limits of examiner reliability

(ii) determination of the grade boundaries, which is judgemental.

In their splendidly pragmatic dissertation, French, Slater, Vassiloglou and Willmott (op. cit.) argue that marking is also judgemental:

We believe that the marks are a quantification of the examiners' judgement of the performance. (p4)

It is probably one of those statements which is more or less true depending on the subject concerned. It is at least arguable that mathematics examinations are fairly objectively marked, whereas English examinations are far more subjective. Such differences are obvious from the marks schemes supplied to each and every examiner. The role of the examiner is therefore seen as fundamental to each of the examination processes and, if the equivalence of their judgement can be trusted at each stage then comparability under social definition must follow.

The fact that there has never been a universally agreed set of grade-related criteria at A-level, makes definitions of comparability based on mastery skills or hurdles difficult to untangle. It can be quite possible for two candidates to exhibit very similar skills on some criteria common to two examinations, but sufficiently differently on the non-common elements to be awarded different grades. In short the domains of assessment may overlap, but the elements not in the union of the domains may dominate the marking. Thus, in occasional paper 1 (1979), the definition that:-

..two candidates who have achieved the same grade in a given subject, regardless of board, mode, or year, should be expected to have attained the same mastery of that subject. (p10)

though a narrow one, may still not be specific enough to use in comparability studies.

Potentially one of the more useful norm-referenced definitions is taken from Good and Cresswell (1988):

Standards of performance.....can....be defined as statistically comparable if equal proportions of the same group of candidates reach the grade in question on each component. (p24)

and was coined to define comparability for differentiated papers within one examination and is particularly applicable to modular schemes. This method of determining grades from a well-defined population is known as 'equi-percentile grading' and is often used to prevent some of the more obvious discrepancies in standards when determining grade boundaries for different components. It also implies that the grade distribution for a given group of candidates will be the same for those subjects which all sat, even if, at the individual level not all candidates received the same grade for each subject. Norm-relating all grading processes can be done, but few involved in the examining process would wish to dispense with the judging element of the awarding process and at some stage allowances have to be made for different entry patterns.

At Advanced level there has never been any attempt to define a minimum standard for a pass or indeed define a modal or median value (grade C perhaps) in terms of the "average" A level candidate. With a large enough population, such as seen with main-stream subjects at GCSE, it may be reasonable to define the median grade in terms of "average ability", such as grade 4 CSE (SSEC, 1963), but A level populations are very selective and differ greatly between subjects. There can be no concept of a norm group at A level and thus no absolute standard. As Christie and Forrest in 1981 make clear:-

Unfortunately there is little consensus and even less systematic research on the crucial differentiating characteristics of populations with regard to examination achievements. (p22)

Since 1981 when this was written, more effort has been put into looking at patterns of subject entry by categorising candidates in terms of centre type and gender but this can only be a proxy for more definitive measures of aptitude or ability. However the development of multi-level modelling techniques (Goldstein, 1995) has enabled research teams (such as ALIS, the Advanced Level Information Service based at Newcastle University which specialises in looking at individual centre's A level results and comparing them to those from other centres while controlling for a range of centre

specific variables) to incorporate other measures of aptitude into research on A level comparability. It turns out that the best predictor of A level results is attainment at GCSE (q.v. Tymms and Fitz-Gibbon, 1991 or Tymms and Vincent, 1994), and hence the differentiating characteristics of the entry population to an A level are best defined by their GCSE results. Until 1994 all such studies had been carried out on a limited, and partially unrepresentative database. Since then several papers have been published (e.g. Goldstein and Thomas, 1996; Gray et al, 1995) which have resulted from analysis using the DFE funded 16/18 database and multi-level modelling techniques. The database is a national one (England, Wales and Northern Ireland) containing matched and merged data from the cohort of 17 year olds taking A levels in summer 1993 together with their GCSE results. Reservations concerning possible bias in the data would not obtain here, and the papers offer considerable insights into the 'value added' performance of schools. Such a database could also be usefully employed to investigate comparability of syllabuses.

The study of comparability issues is not an exact science, but without the regulatory effect that such studies have, the task of selection for employment and higher education would be rendered even more difficult. However, it is not the business of awarding bodies to ensure predictive validity, although syllabus content is often influenced by the requirements of those authorities. More importantly, boards have to ensure the "safety" of their awards, and for this comparability has to be attempted.

Good and Cresswell (1988) define two types of comparability - statistical and judgemental. Though strictly the former was applied to different components of the same examination, most methods of determining comparability fall into the same two categories. Cresswell (1993) has exposed many of the myths surrounding the misapplication of statistical methods and there is no doubt that they should be used with caution. He concludes that:

Crude comparisons of the raw results of examinations...cannot give a reliable indication of their comparability.....and it would be unwise ever to rely exclusively upon statistical analysis of results for information about the comparability of public examinations. (p7)

Judgemental methods are also fraught with problems, as Johnson and Cohen (1983) realise:

...the main inadequacies of the borderline ratification methodologies recently used have been the impossibility of quantifying either the perceived differences in the boards' grading standards or the reliability of these perceptions. (p24)

Maybe the best that can be hoped for is that gross inequalities will be eliminated and that those practices which could give rise to differences can be more closely specified. An attempt has been made by the introduction of a GCE code of practice (SCAA, 1993) to which all boards are signatories.

Comparability Methodologies

Much of the research on comparability within the public examination system has been promulgated either by the Secondary Schools Examinations Council (or one of its later manifestations) or the GCE Boards (latterly including the GCSE Groups). Work carried out under this banner before 1985 has been summarised in two reviews by Bardell, Forrest and Shoesmith (1978) and Forrest and Shoesmith (1985). In his foreword to the earlier of these, A. Robin Davies writes that:-

In a climate of growing public interest in public examinations comparability of grading standards is a popular focus of attention; ... (p5)

Little has changed.

The methodologies used to determine comparability fall into two groups - statistical or judgemental, though it is the former with which this research is primarily concerned.

Statistical

As early as 1970 adjustments were made to the various grade distributions of different boards to allow for the different types of examination centre. Short of the same set of candidates entering the same examination with different boards, this is as close as possible to equating two sets of candidates. This technique has its more modern equivalent in the delta index analysis (a good description of the technique which originated from the SRAC will be found in most recent inter-board/group comparability

studies q.v. e.g. Quinlan or Ratcliffe, 1993). This adjusts each board grade distribution within a given subject to allow for entries of candidates of differing abilities as exemplified by their centre type so that fairer comparisons can be made from the modified distributions.

Some of the methodologies outlined in the reviews may appear quaint to modern eyes, but nevertheless the points raised are still valid. Although duplicate subject entries to different boards is now rendered impossible with the agreed timetabling arrangements, comparability not only depends on having equivalent grade demarcation points, but ranking candidates in the same order. This may seem a trite observation, but actually is fundamental for within subject comparability. It is implicit that the same weight is given to the same skills in a given subject.

Subject pairs analysis when used for comparing the results of the same group of candidates in two different subjects was not entirely convincing, though it was believed that comparative subject pairs when considered board by board might indicate discrepant grading standards. The technique was applied dually as described by exemplification in Forrest and Shoesmith (1985):

(i) If we consider the entire group of candidates taking both Physics and Chemistry, say, in a particular board, how does the distribution of grades in Physics for that group compare with the corresponding distribution in Chemistry?

(ii) If, for example, the average performance in Physics of those also taking Chemistry was half a grade above their average Chemistry performance in eight of the nine boards, but half a grade below in the ninth, we might well infer that standards in that board were out of line with the others in Physics or in Chemistry or in both. (p10)

The tendency of centres to pick and choose boards, together with the decline in those taking certain types of A level, means that it is often impossible to find a sufficient number taking any two comparative subjects with one examining body who could in any way be described as representative of the population of candidates for those subjects.

As Bardell, Forrest and Shoesmith (1978) point out:-

From a statistical point of view, subject pairs analysis is equivalent to the use of a monitor test (p18)

and these were the basis for a number of studies undertaken in the 70s. The problems of bias and relevance as well as the undoubted difficulties involved in their administration has led to a decline in their use. A recent article by Paul Newton (1997) also casts doubt on the subject pairs analysis process. However, any compulsory module within a modular scheme can be considered in the same light as a reference test to which no problems of relevance attach, since it is an integral part of the examination; although bias, certainly in terms of inconsistency, would need to be disproved.

More recent work on comparison of boards and syllabuses at A level has been carried out by CEM at the University of Newcastle at the behest of the GCE Secretaries (Tymms and Vincent, 1994). This used a reference test in a different context - that of multi-level modelling. They were looking for significant differences at both board and syllabus level and rather worryingly concluded of mathematics:

However, the differences in Mathematics were associated with modularity and explanations other than leniency/severity would seem to account for the observed discrepancies. (p1)

The argument they employ is that weaker candidates would choose not to cash in their modules and therefore the examination could be seen as more lenient. It is hard to endorse this explanation except at the grade distribution level. Controlling for ability, using data from GCSE results, a reference test or both, should mean that the grade distribution would be seen in the light of a stronger candidature.

Judgemental

Cross-moderation studies used to determine judgemental comparability fall into two categories (Bardell, Forrest and Shoesmith op.cit.), those of identification when:-

the participants are required to locate a grade borderline from among the scripts in a batch

and ratification in which:

participants are informed that scripts relate to a specific borderline and they are required to decide whether their or not the board's decision was in their judgement appropriate.

(p28)

In addition "scriptless cross-moderation exercises" were prosecuted by comparisons of question papers and mark schemes for differential demand.

Most cross-moderation exercises today are a combination of the latter two exercises; a factor analysis and syllabus review which is rated by each scrutineer for demand, and a ratification exercise. However, the central theme of this thesis is that judgemental comparability is assumed, and the statistical analyses carried out on the basis that grade standards are equal whatever their provenance or imprecision.

It would be inappropriate to finish this rather historical overview without mentioning the work on comparability by Nuttall, Backhouse and Willmott (1974). This may be considered a seminal work on statistical comparability because of the number of different methods of analysis employed. They use "different kinds of tests" which they analyse by regression and the guideline method, subject pairs analysis where mean grades are compared, mean grade in subject compared with mean grade in other subjects attempted and analysis of variance. Until the advent of multi-level modelling using multiple regression techniques in 1986, these five methods were used (in various guises) for almost all statistical comparability studies.

The thrust of this thesis is a statistical investigation of the various facets of comparability. Underlying this is however an assumption of the rigour of the judgements which have already been employed in determining the grades which are awarded i.e. an assumption that there is judgemental comparability. Again this is an implicit use of the type of comparability defined in the social definition, a comparability anchored in the trust accorded by the users of the qualification in those charged with their delivery.

Factors Influencing Comparability

Demand

In line with the two measurement processes (norm and criteria referenced) and the two types of comparability, there are two major factors from which inequalities could arise. The first might be called demand. Pollitt, Hutchinson, Entwhistle and de Luca (1985) suggest that there are three facets to the task of answering questions - subject or concept difficulty, process difficulty and question difficulty. The demands of syllabus, not only in terms of content but in student and teacher motivation, teaching time etc could be labelled subject difficulty, process difficulty is more concerned with what on might call exam technique, the relating of what is familiar to what is unfamiliar and question difficulty relates to factors such as "helpfulness" of the question, whether it leads or is open ended. These three factors are currently used to determine demand of various syllabuses and questions in the syllabus review for cross-moderation exercises. One rather noticeable aspect of these reviews is the tendency for examiners to rate other boards "demand" to be more lenient than their own i.e. ratings are negatively biased (q.v. for example the 1993 GCSE comparability studies in history and geography).

Since 1985 when the review of comparability studies was published, there was little activity in the field of A level comparability (probably due to a concentration of research effort on GCSE which was first fully implemented in 1988) until 1987, when the boards collectively as represented by the members of their Standing Research Advisory Committee (SRAC) embarked on a study of mathematics A level, concentrating in the main on the different routes to a double award. The brief was much wider than usual, and in consequence although many tried and tested methods were used, the range of data which resulted highlighted areas of concern which were perhaps little regarded before.

The syllabus and question paper review, or scriptless cross-moderation which was the basis of study 1, attempted to analyse "demand" i.e. how difficult the papers were. It proved impossible to quantify. The effect of compensation made some routes to a double award in mathematics easier than others, and the same may be found in modular schemes where the flexibility allows different combinations of modules to be aggregated for a subject award. Miscellaneous combinations of papers also created

variable "psychological" demands. These are essentially properties of the structure, and although the researchers found many similarities between the boards, they concluded that:-

the structure of the examination was the single identifiable factor in contributing to any difference in demand. (p43)

This conclusion clearly has implications for modular examinations and is one area which requires consideration.

An interesting aspect which arose from study 3 was the art of question spotting. It would appear that familiarity with the style of some examiners made some questions more accessible. Rather like crossword addicts, practice of past papers helped candidates become adept at recognising a particular examiner's style and answering in a way which would maximise the marks awardable. How this would affect candidates for modular examinations with its heavy reliance on question banking for some modules is unclear.

More recently a report on the SAD project (Pollitt et al, 1998) concluded:

the usefulness of differently structured questions therefore depends on the aim of the exam (p127)

and the control that examiners wish to exert over the outcome space. They observe that structured questions usually elicited better responses, but that did not mean that it was easier to gain a higher grade. However, attempts to make an examination both accessible and demanding reduced the ability of an examination to discriminate well and reduce it reliability. Thus the structuring of questions does not imply, per se, any decline in grading standards since this is allowed for in the awarding process.

Other factors which may affect demand is the students' attitude to different syllabuses within a subject and to their different perceptions engendered by classroom processes. The Tymms and Vincent (1994) report concluded that:

...there was no evidence to suggest that attitudes towards subjects varied across syllabuses, and it is of particular note that none of the project or modular syllabuses formed strong feelings in the students that took them when compared with others. (p15)

This may contrast with teachers' views, but attitudinal effects may be considered to be of little importance when considering comparability. The same report also adds "that there were very few substantial dissimilarities in the way different syllabuses were taught."

To dismiss entirely the inter-action between student and conventional/modular syllabus on the basis of this one report would not be sensible. However perspectives are important, and it is surely the case that other more important differences should be investigated first.

As may be concluded from this discussion, quantifying differences in demand, even within the same subject, is virtually impossible and almost always reduces to a question of judgement. It is a judgement exercised by examiners when setting question papers and mark schemes and for the purposes of this thesis is assumed. Moreover, and fundamentally, it is the business of awarders to ensure that any perceived disparities of demand are considered in the setting of grades. Content, however, is a different issue and this, at least, is amenable to investigation.

Aggregation and Awarding

It is difficult to separate these two processes because however great a part judgement plays in the awarding process, there is always an element of norm-referencing and the mark distributions from which grade distributions emanate are the result of the aggregation process. In study 3 (SRAC, op. cit.) it was felt that the inclusion of double award candidates in the data for the single award could depress the numbers gaining the higher grades because the standard would be set too high. How much more this problem would obtain when grading modules which are strictly stand alone and would include resit candidates as well as those intending to aim for double or even triple awards is incalculable, but, on the basis of this research, it must exist. One conclusion from the study is inescapable:-

there is a manifest discrepancy in grading standards between mathematics and further mathematics. (p48)

If this is translated to the modular situation, it may be expected that there could be a discrepancy between grading standards of different modules, which again has implications for comparability.

It is not the purpose of this study to re-invent the wheel, and Thomson's authoritative report on aggregation of modular schemes investigates methods of combining the marks from the constituents of modular examinations. Although the study was designed round certain MEG GCSE syllabuses, its recommendations have been adopted for A level application. He researches the various ways either module grades or marks can be combined to produce a subject grade. (Aggregation for traditional examinations is well documented in the various codes of practice). He stresses an important requirement of any aggregation procedure:-

...it is vital to ensure that the aggregation system used carries forward as much information the maximum amount of information about the examiners judgements as possible to the final grade. (p51)

The aggregation procedures investigated by Thomson (1992) were:

(a) The 'raw' method - an aggregation of the module raw marks
(b) The points system - each module is awarded a point score equivalent to its grade and these are summed to give a total points score for the syllabus
(c) The arrangement system - the module grades are ordered for each candidate and the grade awarded is laid down depending on an algorithm which for five modules gives a syllabus grade which is the lowest of the top three module grades, provided the last two grades in the series are not too low.
(d) The grade point average - this is the average points score derived from module grades.
(e) The uniform marks system - a method of scaling module marks to a common scale (explained in more detail later).

His recommendation on a theoretical basis was the adoption of the uniform marks scheme, or standardised score. The cross-moderation identification study was designed so that:

Scrutineers' judgements of overall quality were to be compared to the 'system referenced' grades produced by the different methods of aggregation. (p59)

The result was an endorsement of the recommendation to use the uniform marks scheme (or UMS) to grade modular syllabuses based largely on the greater accuracy of such aggregation because much of the original marking information was retained and as a result fewer 'vicarious' grades awarded. Module scores were to be reported, rounding up of 0.5+ and allowing for regression at grade A. (A fuller explanation of the grading schemes for modular syllabuses is given in the next chapter).

In 1994, modular A level schemes used a variety of grading methods, but subsequently all boards adopted the same basic scheme (see appendix A) which essentially takes all Thomson's points except regression. However it is clear that there is far from universal agreement still, much of controversy stems from the way different subjects are marked and how much of the full mark range is used i.e. the true weight of each component. This is specific to grade A at the module level, but has consequences at the subject level because of the variation in compensation that can be achieved by a high score. Another particular problem is that of "premature approximation" which pervades all points-related methods of aggregation (i.e. marks to points to grades). It is becoming well understood and has now been dropped by all boards originally using it.

Discussion

If the literature on comparability is agreed on one maxim, it is that defining comparability is far from simple, and constructing methodologies for discovering its presence (or not) is not straightforward. It is not just that the anchor points, grade boundaries, for the scaling may not represent the same standard even within a subject, but the variations of weight, content and demand impose subtle mutations on seemingly similar regimes. When such regimes are blatantly different, the determination of comparability becomes an order of magnitude more difficult.

There is little in the literature which helps the understanding of the workings of modular schemes. Many of the comparability definitions rely on the 'same candidate' definition. Within modular schemes, the population varies even between modules and such definitions are not appropriate even within the scheme. Orthodox methods of determining comparability, primarily that of cross moderation (Forrest and Shoesmith, 1985) are difficult to apply to cross linear-module comparisons. The fundamental difference between the two schemes is that with linear schemes there is some concept of holism, within the modular environment, the module is a separate and distinct entity and is awarded as such - or should be. Each module may encompass different methods of assessment, and the learning strategies are also likely to be different.

Beginning to appear in the literature are articles centring on the comparability (or otherwise) of modular schemes (Taverner & Wright, 1995; Tymms & Vincent, 1994), but the data on which these are based is somewhat suspect and they are only concerned with generalisations i.e. comparability between two or more syllabuses, and do not focus on what must surely be the first concern in modular comparability, that of 'between module'. Nor do they attempt to answer the question of 'why' might different, or indeed the same, outcomes be expected.

The analysis by Cresswell (1996) highlights the difficulties inherent in using any one definition of comparability, even if it was amenable to analysis. If comparability could be proved under any one of the definitions, it would leave questions about the others. Potentially, therefore, the most obvious way forward is to adopt one notion - the social definition - and examine its facets. This definition invites us to make assumptions about comparability based on the competence of awarders (who are also question, mark scheme and grade boundary setters) and the use to which grades are put. It also implies that the inferences made from two equivalent examinations are the same and thus equally valid. Although questions have been asked and some doubts voiced, there is no evidence that grades from one type of A level are regarded as inferior to any other - whether by subject or assessment scheme. We might start, therefore, by making the assumption of judgmental comparability and equivalent validity, and attempt to investigate its robustness using statistical methods; a theme which is expanded and investigated in later chapters.

CHAPTER 3

It is Ages Ahead of the Fashion - The Modular Context

The rise and rise of modular A levels in recent years culminated in 1996 when for the first time all English A level boards offered at least some modular syllabuses, and all mainstream subjects were available in a modular as well (usually) as a linear form. In mathematics, biology and chemistry there were more modular candidates than there were for linear schemes, with near equivalence for physics (SCAA, 1996). Schools have clearly decided that this type of assessment is, for reasons of flexibility, motivation and feedback (Gray, 1992; Nickson, 1994; Porkess, 1995) one they like. The opportunity to spread the examining over more than one session also adds to its appeal.

It is, however, perhaps inevitable that the impact of modularity should be differential across the different subject areas. The maturity factor allied to rather more parallel teaching across modules in many of the arts subjects, especially English, has led many of these students to choose to sit all their modules terminally. The sciences and, in particular, mathematics, do lend themselves more to this type of assessment and it is no accident that these subjects made up the vanguard of the 'new wave' of modular syllabuses.

Whilst modular examining has met with far from universal approval, its take-up by teaching institutions indicates not only its success as an assessment regime, but also that its impact on curriculum issues is far from unwelcome. The change to distributed assessment marks one of the most substantive modifications to 18+ examinations from the A level of 1951. when the change from school certificate to subject based GCE examinations was implemented. However, the recent popularity of modular A levels is somewhat surprising given that credit accumulation schemes have been around for a long time.

Given both this popularity and the availability, the question arises as to the reasons for the mutation from the tried and very much trusted methods of A level examining throughout the last 45 years to the current, somewhat hybrid schemes. Pressure for change can be both internal and external, but the context for change is primarily dictated by the culture in which it occurs. The assessment culture of the late 20th century is one racked with profound change brought about by political ideology, which although not aimed primarily at 18+ examinations, has created the current climate.

It is probable that the A level examination, for so long the 'gold standard', would be subject to, at least, some modification if only in the range of subjects offered. This is because of the increase in university populations and the diversification of courses provided (partly brought about by the re-classification of the polytechnics in 1993) and the implementation of both GCSE in 1988 and the National Curriculum at Key Stage 4 in 1994 which, together, saw an increase in the number of 16 year olds who were reaching the required standards to embark on an A level course of study. Gone was the traditional highly academic cohort to be replaced by a much larger, more diverse set of students whose needs could not always best be accommodated by the orthodox linear examination in a traditional subject. The examination system needed to change to reflect this.

Examination regimes on the whole are reactionary and the social changes which have taken place over the past 100 years have led to variations in schemes of assessment and to the control which is exercised over those in the public arena. Most can be described as a progression, and even some, innovation.

Some History - Pre 1951

The Chinese are often credited with the devising the first examinations about 1000 B.C. (Ingenkamp, 1977), using them to determine entries to their civil service. By the 7th century A.D these had developed to embody three different forms of assessment - the oral test, T'ieh (an objective test) and T'se (a project question), much in the way that we might combine elements today (Brereton, 1967). These examinations also had a clearly defined purpose, that of selection through competition, and although the class of those who could take the examination was severely limited, it did mean that control of the state was not always in the hands of the same families such as was found in those countries with a strong hereditary ruling class. As such, examinations can be seen as a means of spreading social control rather than limiting it.

Despite such a long history, examinations in Europe and the Western world were considered socially irrelevant (Ingenkamp, 1977) because they were usually in the form of verbal debate, or disputations, between those young men and their tutors who were already of the social élite. The direct descendant of this type of examining today is the viva examination which is sometimes used to determine class of degree (usually for borderline candidates), but is almost always used in the final stage of a doctorate where, as with the

medieval university student, the research student is called upon verbally to defend his or her thesis, though nowadays this will rarely, if ever, be held in Latin.

During the Renaissance, the Jesuits developed a form of written examinations which was used in the European schools (which can probably be traced to the missionary journeying in China of one Matteo Ricci), and hard on the heels of the Jesuitical examining was the appearance of the Abitur in Germany. Later, with the retreat of Napoleon's armies in 1791 came the French Baccalaureat, which, it is suggested (ibid) increased the influence of government control by eliminating subjective judgements (and undoubtedly nepotism) and replacing it with a competition which determined entry to the civil service.

At the same time, written examinations were being introduced at the great universities of Oxford and Cambridge (Broadfoot, 1996). They were used to rank final examination students, and as such could be considered competitive, but in no way did they alter the social status quo, since there was still no mechanism for selecting those who went to university in the first place. This was based on wealth, privilege and, in rare cases, patronage.

However, the European experience together with the opening up of the Civil Service following the Northcote-Trevelyan Commission report of 1853, soon led to the introduction of the examination as a route to advancement in England in the mid 19th century. The British first started using the written examination as an entry qualification to the Indian civil service in 1851 and in England in 1870 (Ingenkamp, op.cit.). Shortly afterwards came the appearance of written examinations in America and the first use of what came to be known as a credit transfer or modular assessment scheme.

The flowering of the examination method of selection cannot be explained without appreciating the background of industrialisation which was becoming socially more pervasive. The increasing needs of empire and industry meant that no longer could the requirement for knowledgeable and skilled professionals be satisfied from the usual sources of aristocracy and privilege. Indeed it was unlikely that many of those traditionally destined for leadership had the necessary skills demanded by the new society. Neither were all of the pool of 'educated middle class' fitted for the role and one method of determining who was best suited was that of the written examination. Not surprisingly this led to a number of schools specialising in the teaching and preparation of candidates for

certain of the examinations, especially those of military academies (Eggleston in Broadfoot ed., 1984). Thus examinations directly affected both the social class structure by broadening the 'upper class' through selection by testing and the school curriculum because certain schools started specialising in preparing boys for such tests.

Not only were the routes to the officer class and the professions widened, so too was the route to higher education. In 1837, sometime before the introduction of civil service examinations, the University of London was established (see Kingdon and Stobart, 1988). Unlike the older universities which young men from the élite entered almost as a rite of passage, London set a 'matriculation examination' which was open to all (including, from 1878, women) and which students had to pass before they could be admitted for the BA degree at either of the, then, two London colleges (UCL or King's). Since the minimum age for this qualification was 16, it soon attracted school, as well as internal, candidates, and thus became the first external school examination, and the University of London the first schools' public examination board. The matriculation requirement of a pass in four or five separate subjects remained broadly the same until 1951 and the establishment of GCE.

The University of London has a remarkable record in the field of public examinations; for it can also claim to have introduced the fore-runner to A level (ibid) by creating a two stage degree examination in the 1860s, with stage one, or intermediate, being taken about two years after matriculation, and, of course, it was the first to introduce modular examinations into its degree courses in the 1960s, albeit not entirely successfully.

The Universities of Oxford (1857) and Cambridge (1858) soon established their own 'local' examinations, of which a fundamental part was the inspection of schools, and gradually, together with London, their influence began to take hold as exemption from other entrance tests was made (e.g. Sandhurst) if a candidate had matriculated. Some time later, in 1874, these same Universities created a joint awarding body (the Oxford and Cambridge Schools Examination Board or OCSEB) for the purpose of providing examinations for sixth form boys intending to go to either of the two Universities concerned. In fact the joint board again had a wider remit than just examinations since it, too, was set up to 'inspect and examine schools, and to grant certificates carrying exemption from Oxford and Cambridge college entrance tests and from first University examinations' (CVCP, 1962). It established a pattern for future examining; firstly it was a school examination i.e. schools

entered candidates as opposed to individuals entering; and secondly it set out explicitly to provide a link between school and University which had happened almost accidentally in the case of the London matriculation.

The link between the Headmasters' Conference (HMC) on which were represented the 'First Grade' or top public schools and at whose instigation OCSEB was created remained strong throughout the life of the Board and although it is unlikely that latterly their examinations were any harder than those of other Boards, for much of its life OCSEB had the reputation for setting the most difficult of the public examinations (Howatt, 1974). Interestingly they also published examination results in the Times, not to universal acclamation. League tables were not welcomed, even in 1875.

What was of equal interest was the beginnings of the influence of examinations on school curricula. There is undoubtedly an interesting debate to be had on the issue of which came first - did what was taught in the public schools of the day dictate the subject matter of the civil service entrance and matriculation examinations or did the needs of the new professions and military academies determine what was taught in schools. Whichever it was (and it was probably both) the late 1800s marked the beginning of the rather symbiotic relationship between schools and public examinations which has continued ever since and probably accounts for many of the strengths and weaknesses of the English education system.

At the turn of the century then there were four awarding bodies following somewhat different paths and in 1903 (a year after the beginning of state secondary school education) they were joined by a fifth, the Northern Universities Joint Matriculation Board set up by the Universities of Manchester, Liverpool, Leeds, Birmingham and Sheffield for much the same purpose as the London Matriculation.

There was by 1911 rather a mixture of 16+ examinations - some were internal (a forerunner of the mode 3 CSE examination), some external, all with somewhat different aims both educationally, as an end in themselves, and as a method for providing a selection mechanism for post-school purposes. Brereton (1944) describes this time as a "period of laissez-faire" in school examinations, with no government control, and the setting up of the 'Consultative Committee on Examinations in Secondary Schools' in 1911 was the means by which the process of unification of school examinations was begun,

although the establishment of the common School Certificate was delayed until 1917, probably because of the intervention of the first world war.

The School Certificate was established as a school leaving examination at 16+ and also soon started to replace the matriculation requirement of most universities. However, increasingly, boys (and the vast majority were boys) were choosing to remain in the sixth form and sit the newly established Higher School Certificate at Principal Standard (Subsidiary was also available) which would also exempt them from the intermediate university examinations. The booklet produced by the CVCP (1962) outlines the examination pathway to a university degree which is illustrated in figure 3.1.

It was still fairly common in 1939 that candidates presenting themselves for university entrance would do so on the basis of school certificate credits or matriculation rather than the Higher School Certificate favoured by many schools. The mixture of examination requirements lasted up to and beyond the establishment of the subject based General Certificate of Education examinations at Ordinary (at 16+) and Advanced Level (at 18+) in 1951.

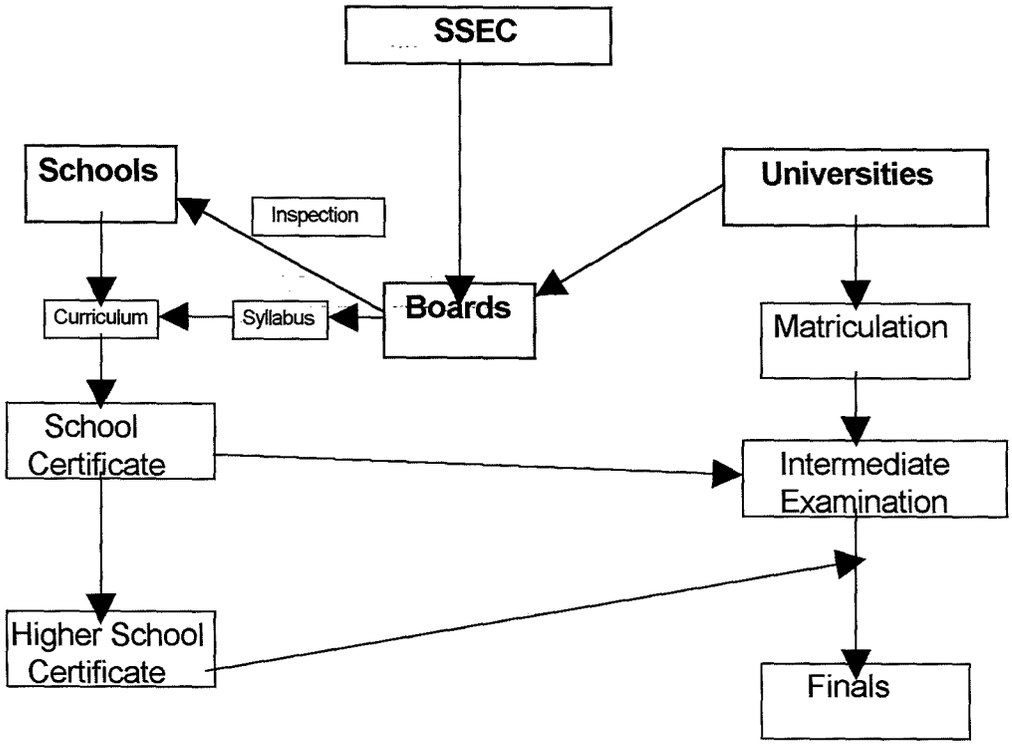
The Control of Public Examinations

The confusion that surrounded the public examination scene was partially alleviated in 1917 when the Board of Education (the second incarnation of what has, since 1839 and until today, in various guises and names, been the government education department) set up an advisory committee known as the Secondary School Examinations Council (SSEC) in order "to secure the necessary equality of standard .....and should perform the function of a co-ordinating authority" (Spens Report, 1938).

The, by now, eight examining boards (Bristol and Durham were very short lived, although the Welsh board is extant) were brought under some control by the establishment of the School Certificate and Higher School Certificate (at 16+ and 18+) and the requirement for formal recognition of their examinations in order that there should be exemption from further entrance examinations to professional training. However, universities still set their own matriculation examinations, and these existed alongside those of the School Certificate as the primary requirement for university. In an attempt to standardise the terminology (presumably) those School Certificate candidates gaining a sufficient number

of credits (marks over 45%) in selected subjects were said to have matriculated. The situation is presented diagrammatically below:

Figure 3.1: The Relationship between Boards, Schools and Universities



There are probably two reasons why the SSEC's influence was not particularly strong. Firstly the government did not always take their advice and this undoubtedly weakened their authority (Brereton, 1944) and they got a fundamental detail wrong, they promulgated the Board of Education's cardinal principle set out in circular 849 of 1918 that "the examination should follow the curriculum not determine it". This latter point turned examining into a bit of a lottery with schools guessing what might be set in the examinations and the boards probably more influenced by university requirements than those of schools. It was never enforceable as a principle, and the failure on the part of the regulatory authority to recognise the influence of examinations on school curricula meant that the Council was unable to impose any effective leadership.

In the early days of public examinations, syllabuses could best be described as 'thin', a very different approach to their construction was taken than is found today. They were probably the result of the deliberations of one or two university dons (Brereton, 1944) with little consideration of what might be practical in a school context. Thus, despite their

growing importance, there was still a remoteness of examinations from most schools not least because only a minority of the school population ever took them. They were definitely considered for the intellectual élite and there was little pressure to alter the situation either from government or its ministries although the examination boards were becoming increasingly unhappy (ibid.). The Spens committee reporting in 1938 were also concerned that "...the School Certificate examination now dominates the work of schools, controlling both the framework and content of the curriculum".

This situation continued until the end of the second world war when the social upheaval it caused affected all aspects of society, including education. The report of the Norwood commission in 1943 was to have far reaching consequences (though perhaps not quite in the ways intended), and, for the first time the links between curriculum and examination syllabus were acknowledged - "Hence, the curriculum is closely linked with the examination" (Norwood, 1943). In fact Norwood recognised three types of curriculum, only one of which included examinations and it was not until the Beloe report of 1960 that it was proposed that the examination population should be extended to include the majority of the 16+ school cohort.

Norwood not only reiterated the need for a tri-partite system of schools based on the three defined types of curriculum (a recommendation embodied in the 1944 Education Act), but recognised the requirement of an examination system to adapt to the rapid changes taking place within the schools. This meant that any change should accommodate not only an expansion in sixth form population, but also should allow a more flexible curriculum which had tended to be firmly rooted in conventional academic subjects. Norwood states:

The examinations have clearly moved much from their original purposes.....the examinations are now a matter of supreme importance to each individual child.....the atmosphere of examinations ...pervading the school at all stages. (p30)

This recognition of the importance of examinations in curriculum matters, which had formerly been unacknowledged, "we find it difficult to accept the dictum that external examinations can follow curriculum", led directly (though by stages) to the control of the curriculum by government agencies which had been kept at bay by their failure to recognise that curriculum could be directly controlled by exerting influence on what was to be examined. However, for the next forty years, until 1982, such control was fairly benign

and allowed room for innovation, made possible in part by the establishment of the single subject GCE examinations in 1951, the direct result of the Norwood report.

Norwood was responsible for the only major change in the 18+ curriculum since the establishment of Higher School Certificate (though not at 16+) although there have been a number of efforts by various governments to impose new regimes, many of which had been firmly rejected by Norwood. Norwood felt the Higher School Certificate to be unfair on a number of counts, not least comparability, and was failing because it was used for the dual purpose of the competitive state scholarship and school leaving examination. He proposed that it should be replaced by a non-competitive qualifying 'school leaving certificate' which would be awarded on the basis of attainment in one or more subjects, and this is the basis of today's subject-based A level examination. The selection for university bursaries was to be by means of a Scholarship examination. In the event this became an adjunct to A level rather than separate from it, and in most cases A level became, and still is, the main qualifier for university entrance.

Norwood never envisaged the change in the tri-partite system of schools with comprehensivisation or the expansion of the sixth form population, and it is instructive to recognise that the single subject approach did prove sufficiently flexible to enable such changes to be easily accommodated. Interestingly the other Norwood proposal of two examination sessions a year (apart from re-takes and a few sitting early in November) was finally fully realised with the coming of modularisation to the A level syllabus. However there continues to be debate as to whether the reform to single subject examination widened the sixth form curriculum (a declared aim) or narrowed it.

The Norwood reforms were carried out under the auspices of the SSEC, which probably accounts for the relatively gentle transition to the new system. For much of the fifties GCE, at both O and A level, was primarily the province of the grammar and independent schools (the 1944 Education Act had originally prohibited secondary modern schools from GCE examinations, (Eggleston, 1984)), but increasingly top streams in secondary modern schools were entered for some O levels, and the raising of the school leaving age to 15 as a result of the 1944 Education Act increased pressure for a school leaving examination for the less able. To oversee these examinations a number of CSE boards were created, but none of them survived the transition to GCSE as individual institutions, all being subsumed by the more prestigious GCE boards.

The Beloe Committee reporting in 1960 saw the effect of the 1944 Act and the inherent unfairness of the somewhat limited GCE examinations. They recognised the educational benefits (Beloe, 1960) of including teachers in the conduct of examinations and the desirability of extending the entirely external nature of board examinations to include teacher input. Suffice it to say that as a direct result of their report, the CSE examination was established in 1963 and included a number of innovations which were eventually included in the GCSE examinations of 1988 and, in some instances, in A level.

The change to include upwards of 60% of the 16+ cohort in taking a public examination of one sort or another also saw a replacement of the SSEC by the teacher dominated Schools Council in 1964. Their report on examining at 16+ two years later makes reference to the need for increased flexibility in the examination system in anticipation of additional examinees on the raising of the school leaving age to 16 in 1972. The considerable overlap of CSE and O level candidates (12.5% entered the same subject in both, Bloomfield et al., 1977) and the recognition of the continuum of ability (Schools Council, 1966) was also beginning to put pressure on the dual examination regime.

Also in 1960, further education colleges were becoming increasingly popular and a seventh GCE examination board, the AEB, was created to serve the needs of these institutions. Within a fairly short space of time they were getting entries from the state school sector and became a major force in the examining world.

The rise of mode 3 CSE examinations allowed, for the first time at this level, innovation in assessment methods, including modular approaches. Because of the plethora of syllabuses which emerged from the CSE method of examining, it is very difficult to pin down those which were modular, although Moon (1988) describes a number of modular schemes which are probably more properly classified as credit accumulation regimes and which dominated the curriculum since, when properly devised, all aspects of the timetable could be accommodated within the modular framework. A number of such schemes of varying complexity were implemented in the 1980s (Moon, 1988).

This does not imply there were no problems. There were considerable practical concerns - how long should a module be; how were credits to be accumulated and there were a number of criticisms outlined by Moon (ibid.) mainly concerning coherence of a whole

curriculum approach and assessment methodologies. Nevertheless, the explicit structure of the schemes, the clearly defined objectives allied to the growing implementation of Records of Achievement and support from vocational bodies such as TVEI, meant that modularity saw the light of day both as a methodology for devising a whole school curriculum, and as a method for assessing achievement at all levels.

Despite an enthusiastic band of supporters, there were probably other reasons which restricted a wider acceptance of the modular approach. One of those was the nature of the government educational watchdog, which, since 1982, was split between the SEC whose responsibility was assessment and the SCDC which oversaw the school curriculum. This explicit attempt to control curriculum matters by the DES was probably a direct outcome from the speech by James Callaghan some years earlier in 1976 when the idea of a 'core curriculum' was first floated. Whatever the reason, by the time a full National Curriculum had been formulated in 1988, it had become clear that there was little future in curriculum reform emanating from other sources. The nascent modular schemes rarely outlived the amalgamation of the CSE and GCE O level examinations in 1988, and those that survived had a very short life thereafter. The SEC which had, together with the SCDC, only existed until the introduction of GCSE, were very unimpressed by modular schemes of assessment even within single subject areas (SEC, 1987) and helped sound the eventual death knell of almost all such schemes at 16+. This may have been partly due to the attempt at modularisation at curriculum level, not just for individual examinations. A few enthusiastic schools were not enough to convince the majority, but the idea of modularity did not entirely die. Neither did the government quangos - they were replaced in 1988 by SEAC (assessment) and NCC (curriculum).

Whilst these massive changes had been taking place at 16+ reflecting, as they did, an almost complete change in philosophy from examinations only for the most able to examinations for all, with the emphasis on positive achievement, little obvious movement was seen at A level. There had been a number of attempts to reconstruct the 18+ curriculum but all were doomed to failure (see Kingdon, chapter 3 (1992) for a full discussion). However, following the Robbins Report in 1963 there was an expansion in Higher Education meaning that there were more university places available. The passport to these places was A level, and so the numbers staying on in the sixth form increased. Gradually the number of subjects on offer also expanded to reflect the requirements of the new cohorts and these were not always in traditional subject areas. But for the vast

majority, A levels were examinations taken once a year with a single form of assessment - the written examination.

It is argued here that one of the reasons for the failure of a number of well intentioned reforms of the sixth form curriculum was the inability of reformers to disentangle curriculum issues from examination issues - even when they recognised the difference. Kingdon (1992) makes the point that dissatisfaction with the curriculum usually manifests itself in proposals for reform of the examination system. Whilst ideally it might be possible to design a curriculum in which examinations had a place, but not one of dominance, in practice the needs of universities for suitably qualified students and of students requiring a place at an Institute of Higher Education have dictated the sixth form curriculum since Higher School Certificate days. In a generally splendid discussion of these issues, a Schools Council Working Paper (1972), when considering another doomed proposal (the CEE at 17+), declares:

We have been at pains throughout this report to stress the pre-eminence of curriculum over examinations and our belief that the former should determine the latter, and not the other way round. (p85)

However, the examination syllabuses which are designed in response, in part, to the needs of Higher Education for suitably prepared students, take up most of the timetable, and the 'voluntary' nature of attendance at education establishments post 16 renders the imposition of a full and balanced curriculum impracticable. Narrow it may be, but if that is what is desired then the curriculum will be wholly determined by the examination syllabuses. It is suggested that had reformers based their new curricula around an examination and assessment structure founded on university requirements they might have had more success. Kingdon (ibid.) also observes that it was the competition for university places which inhibited reform. There is some irony in that Higher School Certificate was changed in order to provide a qualifying examination for university entrance and to dispense with the competitive element (Norwood, 1941), but the philosophy of positive achievement, so wholeheartedly embraced by the GCSE examinations, always took second place at A level. As long as there were more potential students than there were places, university entrance requirements were always going to dominate the sixth form curriculum.

The candidates for A level were undoubtedly influenced by the changes at 16+. The influence firstly of CSE, and latterly by GCSE had affected expectations, not only in the type of assessment with which they would be confronted, but also many more of them assumed that they would proceed to further or higher education post 18 and were looking for more than the traditional, highly academic course, even in traditional subject areas.

Additionally, although the older universities still kept to the equally traditional degree courses, many of the newer institutions employed continuous assessment methods in the awarding of degrees. Some courses were modular in nature and these were felt to help the weaker candidate by splitting assessment into manageable tranches rather than relying on one or two sets of examinations. Many degrees became 'combined subject' degrees which were ideally suited to the modular approach.

Faced with the reforms in assessment at 16+ and degree level, increasingly the time had been ripe for change at the intermediate stage of A level, and after a slow start that change was brought about by the modular revolution. This began in the mid-1980s and was undoubtedly influenced by the implementation of the 1988 Education Reform Act and rejection of the Higginson report which had proposed yet another set of reforms at 18+.

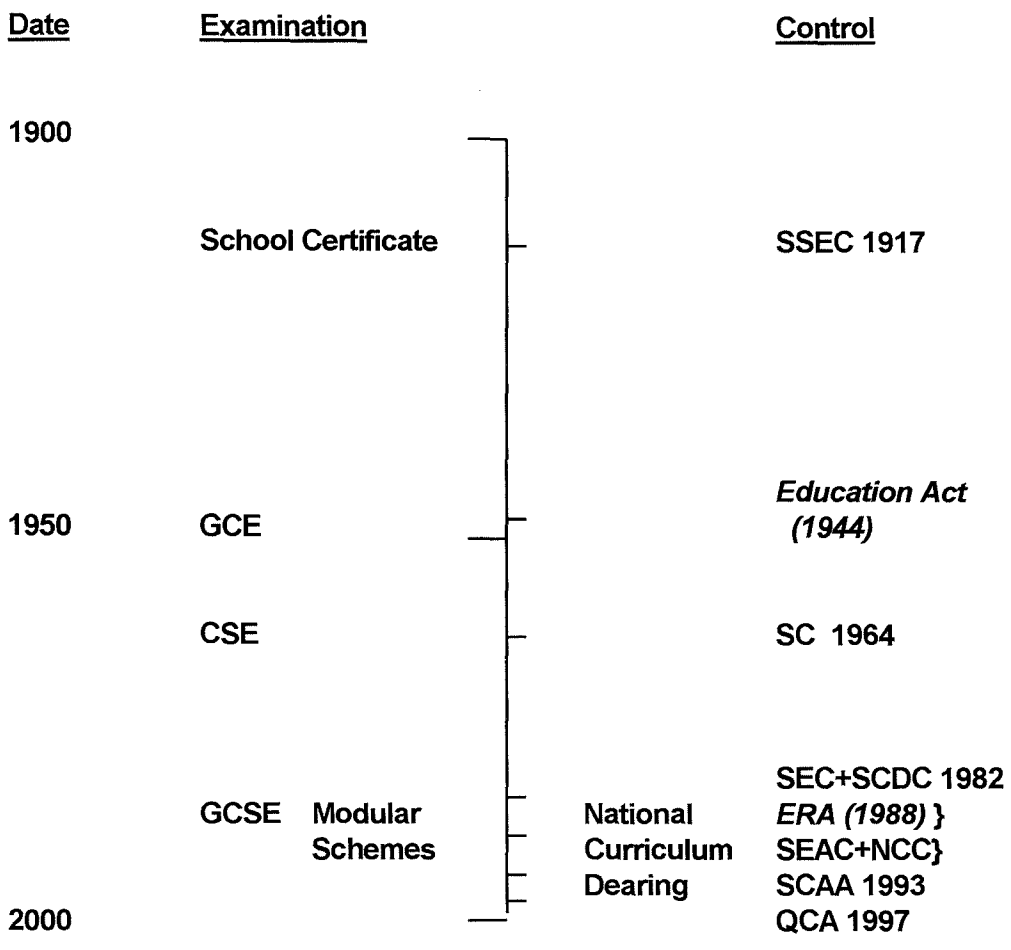
Reform

The diagram below illustrates not only the changes in examinations throughout the twentieth century, but also hints at the government control which culminated in 1988 with the Education Reform Act which gave statutory control of those examinations that occurred within compulsory schooling. This power effectively also controlled examination provision at A level since the regulator has to approve all GCE syllabuses as well as those at GCSE. The difference is subtle; at 16+ the syllabuses must reflect the National Curriculum, at A level, provided subject cores are included, syllabuses are unconstrained - in theory.

That there was a causal relationship between reform at GCSE and changes at A level is moderately obvious, and it is also probable that while government attention was directed towards the imposition of a National Curriculum, and the testing thereof, the impact of modularity at A level was being felt - and liked. It is possible that if the syllabus reformers had turned their attention to the whole curriculum then the movement would have been

stopped - either by the influential universities or by the curriculum regulatory authority. But the piecemeal implementation of modular schemes, first through pilot examinations, then by individual boards and projects, proved successful, and by the time the SEAC and NCC were 're-combined' in 1993 in the creation of SCAA and woke up to the evolution of A level assessment, the momentum generated by the implementation of new schemes was unstoppable.

Figure 3.2: The Pace of Reform



Although there has been no explicit attempt on the part of the regulator to end testing of A levels over a period of time, there have been constraints imposed which have reduced flexibility of the original conception, and, it is suggested, weakened the positive aspects of the schemes. A rigid adherence to 30% terminal assessment has caused technical problems (especially in the case of double mathematics) and the imposition of a terminal synoptic module in an attempt to counter criticisms of fragmentation and incoherence has missed its mark.

The Rise of Modular Examinations

Credit accumulation and transfer schemes originated in the United States in about 1869 (Theodossin, 1986) as their method of implementing the new approach to selection, via examinations, were imported from Europe. Their philosophy was however very different, not surprising since the educationist John Dewey had a hand in their development at Harvard. His ethos of 'self-realisation through customised study' still needed to be quantified. The elective method popular in American universities in the late nineteenth century was policed by a system of credit accumulation. There was also (in theory) transferability of credits between institutions, and certainly between courses. This credit system could also be used to restrict choice if certain courses attracted no credit, so strictly, complete freedom was not possible. The idea of examinations linking one phase of educational experience to another to create a continuum of experience (Brereton, 1944) is entirely consistent with the modular approach to both curriculum and assessment espoused by Dewey who, though disliking examinations, appreciated the need for the motivation that they provided.

The credit system was also in use in Scotland after 1889 when the fixed curriculum in universities was abolished but seems to have limited choice and been fairly short lived. However Scotland did figure largely in proposals to implement modular schemes at school level post 16. The Scottish Action Plan, as it came to be known, was preceded by two proposals for change of the 14-16 curriculum, the result of two reports, one by Munn and one by Dunning (see Scottish Education Department, 1980). These foreshadowed the changes proposed for post compulsory schooling. In the main they were accepted, but the Action Plan itself, proposed a year later in 1983, foundered. Its central proposal of 40 hour modules, known as curriculum components, seemed very similar to the American elective system, with a course of study, composed of a number of modules, tailored to individual requirements. Almost certainly it became too unwieldy, with about 1300 modules proposed, and, in common with other similar pre-GCSE schemes in England, the modular approach was curriculum oriented rather than the assessment driven and tried to offer almost unlimited choice to each student. Hart (1988) suggests that in order to implement the plan there needed to be a complete restructuring of post 16 education which was intended to be inclusive of all educational aims, perhaps overly so. Theodossin (1986) is somewhat more cynical:

The module may arrive with the promise of change and end by accommodating the conventional.

Whatever the cause, the revolutionary changes which would have been required proved too radical and the plan was never implemented.

Theodossin's account (ibid.) tells us that, in England, the first modular courses appeared in the universities. In 1966, the University of London introduced such schemes for its science students, but with little success. Similarly a desultory attempt to do the same was made at Southampton in 1972. In the same year, Oxford Polytechnic also set up its own scheme, with somewhat more success. However, the lead in this area was taken by the Open University where programmes of study and their assessment were based on a system of modules and credits ideally suited to its population. From its inception in the late 60s to today the same system has continued. The system is flexible, allows exemptions and transfers, offers a wide range of options and is not time restricted. Course choice is made on an individual basis within an overarching framework and the options available.

Attempts at modularisation by the vocational bodies in the late 70s and early 80s largely failed to become more widely accepted, probably because there was some confusion between the meaning of modularity and the requirements of a credit transfer system. But following the lead of the OU and Oxford Polytechnic, more polytechnics attempted to modularise some of their courses (run under the auspices of the CNAA), and by 1980 some 25 of those institutions were successfully running some modular schemes. Gradually individualised and credit transfer approaches to degrees became more widespread and gained in popularity both with the providers and the students. The trend away from the all encompassing terminal examination found favour with many and was not unnoticed by school examination innovators.

Prominent amongst these in the school environment were Moon and Hargreaves, and it is they who are credited with starting the modular movement in schools in about 1984 (Jenkins and Walker, 1994) with support from the TVEI. The principles of "a common set of learning outcomes informed by a common set of teaching and learning methods inspired by a common set of learning principles" (ibid.) were the ideals under which modularity was introduced into schools. The top-down approach to curriculum design was intended to allow flexibility and choice within the timetable by means of credit transfers -

assessment was to be built into the learning process which would centre on clearly defined learning and teaching targets. However choice and the lack of course coherence which often arose when free choice was exercised meant that such a dynamic form of curriculum, often engendered by student involvement in the learning process, was thought not to be appropriate to A level although there were plans to extend the TVEI initiative to include the whole of the 14 -19 age range. Government imposed restrictions on coursework and the stress on terminal examinations at the end of the 1980s diluted the aims of the modular curriculum by imposing unwelcome constraints and few schools ever implemented the schemes as originally envisaged.

Project Examinations - and Mathematics

It is necessary to make a slight digression to consider where examinations come from. In the beginning all examinations were set and marked by the boards according to published syllabuses (although schools were free to submit their own syllabuses they hardly ever did so), and this remained immutable for many years (along with most syllabuses). The advent of CSE altered the situation with the mode 3 examinations for which not only did the syllabus originate within a school (or group of schools or LEA) but the assessment (of whatever type) was set and marked by the school, with moderation by the board. This latter process legitimised the examination, but there was little board control. Thus was innovation taken into the public arena in that different forms of assessment combined to make the whole.

The lack of change did not imply that there was no movement for innovation external to the boards. There was, led by the Nuffield Foundation, and they were able, in the late 1950s, through the Nuffield Science Teaching Project, to embody their ideas in a new GCE O level. The breakthrough established an important principle, that responsible and interested groups with innovatory schemes should and would be allowed to enter the public examination arena. As Kingdon (1991) points out, these project schemes often came as complete packages with teaching aids and new examinations alongside the syllabus specification. Kingdon (ibid.) calls them 'curriculum packages' but only in the sense that all aspects of the subject are self contained, not in the sense of influencing any part of the curriculum outside that subject area. This is an important distinction.

One other aspect to the projects was their relationship with the GCE boards. Without the

accreditation of the boards their ideas were toothless since examination syllabuses (arguably) dominated the curriculum, especially at A level. Different projects appealed to different boards, and despite their élitist tradition, OCSEB were in the vanguard of those who were willing to endorse and promote project examinations. This willingness to consider new assessment schemes for inclusion in their portfolio of examinations, accepting that change for worthwhile educational reasons was entirely appropriate, explains much of the reasoning behind the MEI decision to approach OCSEB with their new modular scheme in the mid 1980s.

The first mathematics projects were instituted in part, because of the Bourbakist movement of French 'architectural' mathematicians who, at that time, had a profound effect on the teaching of mathematics through their ideas on structure and formalisation (Moon, 1986). These had led directly to the 'new maths' taught in primary schools in the 1970s. This new maths also motivated innovators in the projects to the setting up of new syllabuses at secondary level. The first of these in 1963, SMP, was the brain child of Professor Bryan Thwaites of Southampton University, and was a collaboration of teachers from a number of leading independent schools who were interested in teaching, and assessing, the new maths. The first syllabus which espoused the new thinking was initially rejected, but eventually the 'Contemporary School Mathematics Project' was accepted by OCSEB (Moon, op. cit.). The Nuffield Foundation, whose mathematical concerns were with the implementation of the new maths in the primary curriculum eventually entered the public examination arena in 1994 with its A level accredited by UODLE.

Not all project examinations were in the field of mathematics and science, though many were sponsored by the Nuffield Foundation. With the change of the advisory body to the Schools Council in 1964, dedicated to encouraging reform, came a proliferation of project examinations, a few of which have survived. The spirit of CSE prevailed, even at A level, and there was a sea change in the philosophy of evaluative assessment. In the words of Kingdon:

The curriculum development projects often created the need for new examination techniques to assess the skills they had encouraged, and this in turn led to the introduction of further project examinations. (p75)

One of the new projects was MEI.

The Mathematics in Education and Industry Schools Project, MEI, was founded by B.T. Bellis whose concerns lay with the mathematical needs of industry. Through his efforts in setting up a consortium of similarly interested teachers, new syllabuses were created with the stated aim:

To promote the links between Education and Industry in Mathematics at Secondary School Level, and to produce relevant examination and teaching syllabuses and support material.

Their first MEI O level examination took place two years after its inception in 1965, with A level following two years later. However, in terms of real innovation, the breakthrough came in 1989 with the acceptance by SEAC of the first specifically designed modular mathematics syllabus, accredited by OCSEB.

The highly structured nature of the 'new maths' arguably made such a syllabus inevitable, but it was at least thirty years after the beginning of the new maths movement that a modular mathematics syllabus found its way into the examination provision.

Another force in the area of mathematics education was that generated by the Cockcroft report of 1982. Cockcroft set out two principles for mathematics examinations; firstly that any evaluative assessment "should enable candidates to demonstrate what they do know" and secondly the "examinations should not undermine the confidence of those who attempt them". One might say that the traditional A level examination conformed to neither of these principles - there is evidence that the course itself was less than motivating and that even students who perform well in assessments still see mathematics as a difficult subject (Bell, Costello and Kuchemann, 1983). The Cockcroft committee also suggested that "assessment needs to be made over an extended period", though in fairness this was strictly applied to coursework. Cockcroft did not have in mind modular schemes of assessment when he presented his report, but from the tenor of his words one suspects he would have approved.

Of somewhat more relevance was the recognition that the mathematics A level population had changed from those who, in the first years of GCE, were predominantly scientists to a more heterogeneous mix of abilities and this population required a course which was

"balanced and coherent in its own right". For this reason Cockcroft was keen to see an element of applied mathematics in A level courses, but he recognised the "difficult and intractable problem" of providing the mix of requirements to satisfy all students within one syllabus. This problem was solved when modular schemes became available and most schools were able to offer choice within a single course, at least sufficient to suit most needs.

MEI was not the first modular A level syllabus and would not claim to be so, although it was the first in a mainstream subject. Modularity had been rumbling on as a method whereby innovation might be achieved at 18+ for several years. It has been noted by several authors (e.g. Kingdon, 1991; Matthews and Leece, 1986) that change at A level has been evolutionary rather than revolutionary. In other words changes were gradual, and only occurred when conditions were propitious; despite many attempts, change was not enforced. However, the agent for change was never the government, despite their criticism of the system and attempts at reform, or even the boards; examination and assessment innovation in most cases originated from the projects.

Modular Curriculum and Modular Assessment

Curriculum change is the result of a complex series of inter-related factors, social, cultural and educational. Norwood defined three types of curriculum depending on the destination of the student - academic, occupational and general. The recognition by Beloe (1960) nearly 20 years later that students do not fall into nice pre-defined categories set the scene for changes which eventually resulted in GCSE and the National Curriculum.

Curricula have, in the past, often been defined in terms of aims (e.g. DES, 1980; Beecher and Maclure, 1978) with the individual schools being left to deliver those aims as best they can. However, as examinations figure ever more largely in pupils' lives, curricula have been increasingly assessment driven. The ERA of 1988 gave statutory control of examinations to SEAC and started the process of adoption of a National Curriculum at 16+, and a firmer control of examinations at 18+. The Education Act of 1997 strengthens that control. Curriculum development has therefore become a by-product of syllabus development even at 16+ and the attitude of the regulatory authority, increasingly an arm of government, is fundamental to those changes.

Nuttall (in Broadfoot ed., 1984) was not alone in voicing suspicion of the boards' role in devising new curricula, although he praised the introduction of common cores at A level as providing an element of rationalisation (this was not always the opinion of syllabus developers who were often constrained by the need to include common cores into their syllabuses). There is evidence (Reid, 1972) that external pressures on the sixth form curriculum come primarily from the universities of which the examination boards are perceived as an arm. It is arguable as to whether examination boards are radical or reactionary, and which should they be, but the literature is clear that examinations have a central and defining role in post-compulsory education.

The meaning of curriculum has become so diverse that in discussing modularity it is necessary to recognise the three contexts in which it arises:

(i) the school curriculum
(ii) the subject curriculum
(iii) subject assessment and evaluation of attainment.

Holistic curriculum change has proved problematic in its implementation, as evidenced by the imposition of a National Curriculum at 16+. Such curricula are far more deterministic than those relying on a series of targets. Organising the school curriculum on a modular basis requires considerable structural change from current practice, and despite the TVEI initiatives and enthusiasm, the difficulty in providing a balanced and workable modular school curriculum which catered for all needs was, in practice, almost insurmountable. This, too, was the Scottish experience. Part of the problem was provision for the sheer diversity of needs both of teaching and assessment. There could be no concept of 'standard' (so important in reporting school achievement) because of the lack of coherence between and within modules, which may (or may not) contain formal examinations. Standardising credits became an intractable problem and the eventual, much heralded, arrival of the National Curriculum, supported by legislation, meant that schemes, such as described in case study E in Moon (1988) were bound to fail.

When GCSE was implemented in 1988, a number of CSE initiatives were incorporated into the new examinations. Prominent among these was coursework and several of the TVEI sponsored modular schemes. These were described as 'cross-curricular', but only in the sense that the schemes embraced a number of subject areas, as did the modules to

be aggregated, (e.g. applied technology which could include modules as diverse as textile design and microelectronics) and not in the sense that the school curriculum was centred around such choices. The idea was that each module was self contained both in teaching and assessment and thus the subject curriculum was fairly defined as modular.

However, modular assessment is both broader in definition and narrower in concept. It is possible to have a modularly assessed A level, with self-contained assessments taking place over a number of sessions, without necessarily being attached to a modular curriculum. It would be counter-productive to attempt to assess a modular subject curriculum other than elementally. The fundamental dichotomy lies with the structure of the programmes of study. It is possible to define these on a holistic basis (the storybook approach of the Salters' science examinations is a good example) but separate the assessments into discrete elements. The fact that the facility to sit an examination modularly exists, does not necessarily mean the subject curriculum is suited to such assessment. It is of no surprise to note that many A level English modular schemes attract a majority of entry from those candidates sitting all modules terminally. Not only is maturity a factor, but there is sometimes a lack of a clear skills and knowledge differential between the various English modules.

The changing face of assessment initiated by CSE and the broader educationist model (Broadfoot, 1988) was instrumental in bringing about changes at A level. However, there was no suggestion that the sixth form curriculum itself should become modular. The timing of the first A level modular schemes is interesting since the date of implementation seems to be about 1987, a year after the first GCSE courses were started (to be examined in 1988). The 1987 pilot Ridgeway scheme, based on London Board syllabuses, seems to have been a school level modularisation of the A level curriculum following the same TVEI initiatives that figured largely in the 16+ curriculum and embraced the educationist principle of assessment. There was also, in the same year, the Wessex A level pilot scheme devised by a group of LEAs and accredited by the AEB. However, the idea that existing syllabuses can be sub-divided to create modules (used again by the London Boards with its first modular mathematics syllabuses in 1993) were not a particularly successful vehicle for mixed schemes of assessment.

There are a number of reasons for this. Firstly, it is important in successful syllabus design to consider the scheme of assessment ab initio. To marry an inappropriate assessment

scheme to a defined programme of study (for that is what current A level syllabuses are) rarely works well. Secondly, little thought had been given to the aggregation of module scores and the points system employed was unfair. Thirdly, modular schemes must have a pre-defined structure and that too was missing. That structure should ensure comparability of modules, certainly comparability in terms of teaching time and assessment otherwise choice was not possible. There was also a problem of within module coherence and lack of a clearly defined educational aim for each one.

At the same time UCLES had started its Module Bank System, a specifically designed approach in that each of the syllabuses included had been written with a definite assessment scheme in mind (Nickson, 1994). Each syllabus was based on six modules with a pre-defined aggregation procedure. The drawback was that none of the syllabuses offered (from 1987 Art & Design and Performing Arts and from 1988 Business Studies etc) was amongst the most popular, possibly because their future was uncertain. However the pattern was established, and in 1990 MEI presented its 'Structured Mathematics Syllabus' for certification in 1992. This again was based on six, equally weighted, independent modules and the syllabus, incorporating programmes of study in all but name, was written round the assessment scheme. It was possible within the same overarching framework to incorporate different assessment methodologies without any loss of coherence. (A list of the different modules on offer in 1992 is found in Appendix C.) There was a clear structure (following the Bourbakists) and it proved an immediate success.

The last attempt by the government to reform the A level system of examining (until Dearing) resulted, in 1988, in the Higginson report. This too was rejected, but the report contained a number of pointers to contemporary thinking on the future of 18+ assessment. A number of extracts are quoted below:

3.12 ....this involves providing students with clearer targets,.....and the provision of better feedback from A levels into teaching and learning on written, end-of course examinations.
5.5 We cannot emphasise too strongly that understanding should be a dominant theme.
5.15 It would be possible to divide subjects and syllabuses into units or modules.....
6.1 We have argued earlier in favour of assessment at more than one point in the course and in more than one form
6.15 moves...towards forms of profile reporting

Whilst Higginson (1989) was a long way from recommending modularity for all, a number of the most important issues indicated above are addressed by modular assessment and reporting.

By 1989, the time for A level innovation was ripe, but the Higginson proposals were rejected. Public, media and probably government attention was diverted by GCSE and the impending National Curriculum. In their consultation document of 1989, SEAC were still referring to discussion about modular approaches and the need for syllabus rationalisation without apparently realising the vehicle for change was already moving. Arguably, the modular syllabus enshrines the two philosophies so much a bone of contention for so long - that of assessment driven curriculum or of a curriculum driven assessment.

Discussion

There is no clear pattern discernible in the development of A level syllabuses. On the contrary, the picture that emerges of the 18+ examination provision is one of a number of failed attempts at innovation. The assessment revolution has been at the school-leaving age of 16 and it is perhaps inevitable that there have been some knock-on effects at A level. The increase in the number of assessment instruments is just one and a wider diversity of subject provision another. Demographic changes have also ensured a broadening of the A level curriculum. Together these may in part explain the need for more flexibility, a flexibility which has been delivered via the medium of modular assessment.

At the time of writing there is a feeling that modular schemes may have failed to develop in the way that was initially envisaged by at least some of their proponents. Of the seven weakness in A levels identified in the 1989 document (ibid.), at least six were addressed by the introduction of modular A levels. Additionally, as concern has grown over the fate of vocational assessments at least one scheme (not surprisingly MEI) was attempting to address that very issue. Incorporating vocational modules within a modular A level would immediately bring the parity of esteem so desired by the vocational bodies. The syllabuses are immensely flexible since it is easy to introduce new modules and bring more choice to existing schemes. However, the vocal minority has decided that modular schemes are 'easier' than conventional schemes. The basis for such opinion is tenuous, and there is a growing body of evidence (not least from SCAA which had replaced SEAC and the NCC

in 1993) that, in fact, the reverse is the case.

Wholesale change (in the form of the Dearing syllabuses) is again proposed. Unfortunately these schemes seem designed to suppress innovation in the cause of 'rationalisation' and the creation of an educational leviathan which is neither reactive nor radical. It is difficult to see any clear educational aims (apart from government policy) other than the recognition that more of the 16+ cohort are remaining in education and that the curricula on offer need to accommodate the increase. Less and less is GCSE regarded as a school leaving examination, and there are even moves afoot to effect its demise. This would bring A levels into even higher focus and the need for the capacity to incorporate a greater range of ability in their provision seems inevitable. How this can be achieved under the present proposals is unclear, since flexibility is not on offer.

The original MEI conception of credit transfer of modules to other qualifications, specifically those awarded by the vocational bodies has been lost in the attempt to impose tight constraints on what can be offered at A level. The future 18+ provision is currently under debate, but there is little doubt that most of the ideas could be incorporated into a modular curriculum framework and the evaluation delivered by end of module tests. Whether this would be politically acceptable is another issue.

CHAPTER 4

Enveloped in Absolute Mystery - The Search for Comparability

There is little doubt from the literature and from the little feedback that exists from teachers that modular schemes are generally felt to allow an easier passage to qualification than conventional forms of assessment and testing. This is expressed under different guises - "candidates are gaining higher grades than expected" (Moon, 1988), or perhaps, "the weaker candidates are finding the subject easier" (Gray, 1992). Therefore it is legitimate to ask does "simpler", "easier to get high grades" etc invalidate the qualification as not expecting the required standard, namely the same standard as expected from other schemes, or are these expressions shorthand for "user friendly". The enhanced motivation which derives from clearly defined learning objectives, short-term targets and regular feedback (Munky et al 1989) may be one reason for the perception that modular examinations be thought to provide an easy option.

As may be the case with generalisations, they contain more than a grain of truth, even if it is unintended. Although 'easier' is often used as a pejorative descriptor, taken literally, it actually means, inter alia, "freedom from constraint, facility" (OED). Almost all modular schemes do allow a flexibility of choice and assessment not open to conventional methods of examining, and the regime of allowing the resitting of modules within the course induces a facility with the earlier learning objectives (especially in weaker candidates) impossible with a single terminal examination (albeit in two or three parts).

Thus, one pragmatic argument advanced for the introduction of modular schemes centres on their perceived user-friendliness. There has been, in accordance with government policy, an appreciable increase in the proportion of 17/18 year olds taking A levels, from about 18% of the age cohort in 1985 to about 30% today. But within this increase, it is clear that there has been a trend away from traditional subject areas especially in science and mathematics. Rightly or wrongly they are regarded as "more difficult" (q.v. Tymms and Vincent, 1994). If students prefer other, more applied, subjects at 18, it is unlikely that they will ever return to the more traditional areas of study. It is therefore arguable but defensible that introducing schemes for traditional subject areas which are more motivational is consistent with political aims.

Also the increasing "modularisation" at A level is consistent with current trends in higher education, especially at the new universities many of which have been running modular schemes for several years. Sandwiched as it is between GCSEs (some of which may still be considered to retain elements of modularity) and higher education qualifications, it is inevitable that there is and will be an expansion in the availability of modular A levels.

Pragmatism thus ensures that such schemes, embracing the modernist philosophy of assessment, are welcomed. But, it is neither this philosophy which ratifies the qualification, nor the comparability with other examinations, it is their social use. Whilst modular schemes are accorded equal status with linear schemes their 'social comparability' is assured. However, this is not quite enough, and there is an underlying uneasiness reflecting a suspicion of modular standards.

Superficially, there appear to be far more similarities than dissimilarities between modular and non-modular schemes. They are both examination regimes for which curricular guidelines are published. They both result in the award of a syllabus grade and the uses to which those grades are put - university entrance, job applications and so on - are identical. There is no reason to believe that there is any fundamental difference in quality or quantity of subject coverage which is pre-ordained by the methods of assessment - of which there is increasingly more variety. Inclusion of coursework is becoming more popular with traditional assessment schemes, and practical and oral work have always played an important role in the examining of the relevant subjects.

However, it is also argued that the two types of examination have not been developed with the same purpose in mind. To quote from the MEI Structured Mathematics syllabus:-

A guiding principle of this scheme is that each component (module)[1] is assessed in a manner appropriate to its content.

---

1 The author's addition. For consistency the elements which together constitute a non-modular examination will be referred to as components. Those which together form the whole of a modular examination will be designated as modules.

This may be the premier consideration governing curriculum development both now and in the future, but was not a variable when GCE examinations were first introduced in 1951. This earlier pattern is the one which still prevails, although some elements of coursework are gradually being introduced.

The "paradigm shift, from psychometrics to a broader model of educational assessment" (Gipps, 1994) can only be held up as a partial explanation for any perceived differences. Despite attempts to introduce a more educationist approach, the prime method of determining attainment at A-level is by examination either terminally, or both terminally and throughout the course and this must embrace the psychometric principles of true score and reliability. Although profiling is part of the modular outcome, it is more a by-product than an integral part.

Therefore we have two schemes of examining within similar curriculum driven regimes, which, if the critics are to be believed, may lead to different outcomes. It begs the question - why? Are there factors within the modular scheme which would lead to a candidate obtaining a greater "command of substantive knowledge" (Ebel, 1965) than if (s)he had followed a traditional course?

It is possible to view the same question from a different perspective. Why might one scheme of assessment lead to greater evidence of attainment than another? What factors might one take into account in an attempt to explain any disparity? Chapter 1 introduced the two ideas central to this thesis.

The first is the assumption that judgemental grading standards are comparable, an assumption implicit in the Cresswell social definition, or description, of comparability. This assumption is not unreasonable on purely practical grounds for the schemes considered here. Many of the awarders are common to both the linear and modular syllabuses and it might be expected that their internalised standard of what constitutes a boundary level of attainment should be the same for both schemes. Also some of the mechanistic differences between linear and modular schemes in aggregation have either been eliminated (there is no regression allowance in either) or minimised (the UMS scheme is designed to reduce conversion differences).

The second is that there are threats to equivalent validity implicit in the two types of assessment. The first, construct under-representation, concerns the tasks which confront the candidate. This is largely content based and its existence, or otherwise, is amenable to detailed analysis of the question papers and syllabuses. Construct-irrelevant variance is strictly about the number of variables a test measures. However, the effect of construct-irrelevant easiness would be to enhance a candidate's score, whereas construct-irrelevant difficulty might depress it. The factors which might do this are here defined as legitimate and illegitimate variabilities. It is important to emphasise that these variabilities are centred in the assessment and not in the candidate, although were one to affect a subset of the candidates differentially then it could be defined as systemic bias.

Legitimate variabilities are acceptable reasons for enhanced performance, though they are not all quantifiable. Whilst it may be impossible to put a value on motivational factors, it is possible to quantify the effect of, for example, resits. Both of these would be considered legitimate reasons for enhanced performance, even if not always liked. Conversely, illegitimate variablilities are those factors which awarders should be taking into account when they make their grading decisions. Prime amongst these would be question or paper demand. What is clear (Pollitt et al, 1998) is that by themselves differences in question demand do not compromise grading standards and, moreover, there are features of each question type which may affect the level of response. However, variations in question demand between syllabuses explain the reason why awards are made and reliance is not placed on raw scores which are meaningless unless contextualised. The presence of both types of variability, legitimate and illegitimate, explain why there may be differences in grade distributions between modules or modules and syllabuses, although only the illegitimate type will compromise comparability. Determining which of the perceivable variables are legitimate, and which quantifiable, or both, is the concern of this chapter, as is their relationship with the concept of 'comparability'.

The Three Domains

It is relatively easy to list obvious differences in the assessment rationales of modular and non-modular schemes:-

a. The number of assessment opportunities
b. Module choice
c. Maturity
d. The facility, or otherwise, for resitting within the course
e. Performance profiling and feedback
f. Differential module difficulty
g. Population
h. Length and number of examinations
i. Structure of questions
j. More examiners
k. Effects (including compensation) of different aggregation methods

but more difficult to explain why, singly or together, they might justify different results.

The list above can be divided into two sections. Variablities listed under a to e could conceivably lead to legitimately enhanced performances by candidates of ostensibly similar abilities who follow separate assessment regimes. The effect of variabilities f to k should be eliminated as sources of difference during the awarding process, though, of course, this may not always be the case.

There is an additional variablility which may appear to be both legitimate and illegitimate, and that pertains to the syllabus content. Apart from core material, the syllabus developers are free to choose whatever material should be covered in an A level course of study. Options, or modules, extend this flexibility. Provided grade standards are maintained, the differences in content remain a legitimate source of variability. However, it sometimes appears that some material is very much more accessible than an alternative offered by the same syllabus; a difference that sometimes proves impossible to reconcile at either the question setting stage or award time. This will give rise to an illegitimate source of variability.

Another perspective on the same issue would be to consider an idealised situation of two identical groups of candidates, one of which took the linear scheme and one of which took the modular scheme. If grading standards were identical i.e. there was judgemental comparability, then the expected outcomes from each scheme, the grade distributions, would also be identical. However if there were differences, perturbations, to

the distributions, could they be explained in ways which would imply that grading standards had still been maintained? What follows is a consideration of those factors which may perturb the distribution, and whether they may be thought of as doing so legitimately.

The nature of assessment in the educational context varies, but it almost always has an evaluative function. This evaluation has to have a point, or points, of reference, and it is these which are in the public domain. They are universally known throughout the public examination system as grades. It is this grade by which individuals are judged, by which institutions and teachers are judged and by which the evaluators are judged. The public examination system is predicated on the belief that such measures are valid and reliable. As an initial position, it is also assumed that any two equal grades represent equivalent levels of attainment and the examinations and curricula which generated them are comparable.

Lord and Novick (1968) describe three postulates for psychometric measurement:

(i) identify the object being measured e.g. the examination candidate
(ii) identify the property or behaviour being measured e.g. mathematical ability
(iii) identify a numerical assignment rule e.g. marks.

There is no problem with the first of these postulates, but (ii) and (iii) are far from simple. Not included explicitly, though undoubtedly important in the modern educational environment, is the need to identify the method of measurement. It is therefore argued that for measurements deriving from the educationist model of assessment, the postulates, as they apply at the candidate level, can be re-cast in the following way:

(A) identify the domain of behaviour
(B) identify the domain of assessment
(C) identify the domain of measurement.

The three domains are inter-related, but within each of these domains lie differences between the two approaches to examining investigated here and to the variabilities defined above.

Domain of Behaviour

The attempt to define the property being measured points up a key difference between modular and non-modular schemes. It is sufficiently fundamental to allow a tentative definition of modularity to be made.

> ***A modular syllabus is one which consists of, inter alia, separate behaviour and assessment domains for each module.***

This seems fairly self evident, but it is not true of "quasi-modular" or traditional forms of A levels. The contrast is between taking an integrated syllabus and artificially dividing it into components for examining purposes, i.e. a decontructionist model and writing a syllabus in separate self-contained units (albeit mandatory within the whole) each with its own assessment regime i.e a reductionist model.

There are undoubtedly a number of quasi-modular schemes which are set at A level. The approach has been to write the syllabus in the conventional way, and then split up the examinations into a number of components which may be sat at intervals throughout the course. These are poor imitations of the true modular schemes lacking both the flexibility and the advantages of appropriate assessment methods. It is known that for many of these schemes, a sizeable proportion of the candidature still view them as orthodox and take the examinations in one sitting. This is rare with the schemes which embody, rather than subscribe to, the modular ethos.

It is argued that a modular scheme should have a syllabus which is written with the assessment scheme in mind, and that each module should not only be self-contained but also should be of equal weighting to each other. If modularity means anything it means flexibility and the eventual possibility of transfer of modules between syllabuses. Thus equal length modules - in terms of teaching, assessment and weighting - must prevail. Differentially weighted modules are endemic to those hybrid schemes developed to satisfy the call for modularity without seriously embracing their ideology. Above all, a modular scheme must be balanced and coherent.

It is asserted that most, if not all, the differences listed at the beginning of this section are a natural consequence of the reductionist, rather than deconstructionist, approach to

the curriculum. That each may contribute a different reason why one should expect the outcome of a modular scheme to differ in some fundamental way from that usually ascribed, has yet to be argued.

The domain of behaviour, "the property is being measured", is largely circumscribed by the syllabus through the aims, the assessment objectives and subject content. It is recognised that it is quite legitimate to have different syllabus contents without changing the essential nature of the assessment, at least within a given and recognisable subject area. There are numbers of syllabuses with many options, history is an obvious example, where there are clear and distinct differences in the syllabus content for each option, but there is little public debate about the possibility of any one option being harder than any other (though it is a concern of examination boards to ensure within syllabus comparability).

Appeal to the OCSEB certificated modular option of MEI Structured Mathematics shows that in the specific areas of aims and assessment objectives there is no real difference between the modular and non-modular schemes. (This is actually true for virtually all syllabuses where there are linear and modular versions). Even the content, though fragmented by the modular approach, in total is very similar. In other words, what candidates are expected to know and achieve by following a modular course of study is very much what they have always been expected to achieve. The domain of behaviour as specified by the syllabus for modular A levels is *not* where any deficiencies, if there are any, lie. It is *how* they achieve which is so different i.e. the curriculum which such schemes impose.

There are, however, two aspects to this domain. The first is the one described above. The second is far more ephemeral. It is argued that if the domains of behaviour (as defined here) for any two syllabuses are the same, then any disparities between them, including those as diverse as modular and non-modular schemes, must reduce to those of validity. This may be too simplistic, not least because the domain of behaviour is variable. It is, after all, the interaction between a candidate and the learning environment that determines the final outcome. The very act of evaluating some specified characteristic may be enough to alter its nature. That this should not happen is one of the precepts on which the public examination system has, until recently, been based. But if such change is inescapable then the examination system should recognise it and

build on it. This is what modular schemes try to do. Whereas the behaviour or performance, consequent upon the learning process, is, in conventional schemes, constant throughout the period of terminal evaluation: the nature of modular assessment is such that performance parameters change throughout the course as a direct result of the in-course assessment opportunities open to the candidates. Thus it is not just the fact that there are a number of assessment opportunities during a modular course which concerns us, but that their effect is to cause changes which may ultimately be manifest in enhanced performances. These opportunities constitute a legitimate variability and one which must be addressed if expected outcomes from conventional and modular schemes are to be compared.

In summary, the domain of behaviour for a conventional scheme is an integrated whole, artificially divided for assessment purposes which are almost always by terminal examination. For modular schemes, it is argued, there are a number of quite separate domains, one for each module. The union of those domains may be holistic, because each may be assessed within the course of study and the relationship between the candidate and each module domain may change such that a better than expected overall level of attainment is reached.

Domain of Assessment

The distinction between this domain and that of behaviour is largely artificial and is not made to disguise their indivisibility. After all, the fourth in the quartet of syllabus constituents is that detailing the scheme of assessment. Within the educational world there are, it seems, an ever increasing number of different methods of assessment - summative, performance based, dynamic, criterion based, teacher, formative, self - to name but a few. Not all are distinct and not all are used for examination purposes, but it is necessary to make a slight digression to consider those that are. The definitions given here are very specific and strictly are the only ones to apply within the framework of the public examination system.

Summative Assessment

There is probably more agreement over this type of assessment than any other. Until recent innovations (primarily with the introduction of CSE and later GCSE) it was the

only type of assessment used in the public examination system, and is still predominant. It is an assessment which takes place at the end of a course and is used to provide information about the candidate's attainment in that area of study. It always takes the form of a terminal examination and provides no feed-back within the lifetime of the course of study.

Formative Assessment

There is a wealth of different definitions of this type of assessment, and even more misuses of it since formative is often used to describe anything which is not summative. According to the OCSEB Modern Languages GCE Examination Syllabuses 1994 and 1995, it is merely a synonym for continuous assessment. Gipps (1994) states that it is "used essentially to feed-back into the teaching/learning process", but it is not clear whether it is either teaching or learning or both since it has been applied in all three contexts. Moon and Mayes (1994) extend this definition but in ways which are generally not applicable to examination regimes; for example disaggregation and criterion referencing. The TGAT report finds that formative assessment is closely related to diagnostic assessment, and indeed if any assessment provides feed-back it is difficult to see how it would not be used for diagnostic purposes, even if only of the grossest kinds. The difference may be more semantic than real being a descriptor of the use to which the assessment will be put rather than a differential description of the actual assessment.

Wiliam and Black (1995) also put the emphasis on the feed-back provided by formative assessment and its effect on future performance. It is thus particularly applicable to modular schemes since undoubtedly there is feed-back to the student via a result for early taken modules. The effect of this feedback may be to cause the student to resit the examination in an attempt to improve his/her performance and reduce the gap between performance and expectation.

For the purposes of this thesis, formative assessment is defined as any assessment (continuous or otherwise) which takes place *within* a course of study, the results of which are known to the candidate. It may be used for diagnostic purposes.

Ipsative Assessment

Gipps (1994) defines this as assessment "in which the pupil evaluates his/her performance against his/her previous performance." This is a slight variation on the 'attainment of personal target' definition, but is particularly applicable to modular schemes so will be adopted here.

Different Assessment Opportunities

The types of assessment described above are all integral to the assessment structure of modular schemes. Each module is distinct, and when a student finishes studying a module, his/her attainment has to be assessed. Although there must be opportunities for poor attainment in the early modules to be improved, and hence resits, in essence once the module assessment is complete, the candidate moves on to the next course of study. However the usefulness of the feed-back provided by such early assessments should not be underestimated. Only those modules taken at the end of the A-level period of study lack assessment opportunities within the course, (except for coursework itself) and it is noticeable that these generally produce lower levels of attainment.

Module Choice

There are two aspects to module choice. The first is not unrelated to option choice in traditional schemes. The restricted nature of a module syllabus confines the domain of behaviour and focuses knowledge and understanding such that module examinations may seem easier than those for options, but essentially there is little difference. But the second aspect of module choice concerns the time at which the module assessment takes place. If the modules are so chosen, then there are opportunities for early assessments, feed-back and re-sitting if necessary. This may advantage the candidate.

Although many conventional examinations may contain question and option choice which appear equivalent to module choice there is one very important aspect in which they differ - a bad question within a paper counts as does a terminal option paper, a low scoring module may be discounted or re-sat or even, as a last resort, lead to the candidates refusing syllabus certification. Modular schemes, in general, allow little question choice, possibly because individual module syllabuses are fairly narrow,

possibly because there is usually a fair degree of choice between modules and possibly because question choice can only lead to more inequity within schemes which are far from homogeneous in design. It is also not uncommon for candidates to sit more modules than are needed and to discard the worst (provided it is not mandatory and no rules of dependence are contradicted). This is a luxury not open to the traditional examination candidates, but one which could easily boost performances. (This aspect of modular examinations is fully considered in the next chapter).

Differential Module Difficulty

This is more profound than might initially be inferred and is more complex than the issue of choice from which it originates because of the grading process. Although module choice allows different behaviours to be quantified within an examination regime, if measurement could be calibrated against to some absolute scale, then there might (arguably) be a case for assuming comparability. There is no absolute scale and the different emphases at awarding between modular and non-modular schemes are such that inequities may arise. Even choice of questions within components introduces an unintentional element of discrimination and has been the subject of many investigations (summarised by Wood 1991), but not only may modules suffer from this, they are also subjected to a grading process. It is not uncommon for mathematics candidates to gain high grades in early modules and lower grades for those sat terminally. Determining modular grades is far from straightforward, and it is not surprising that awarders may resort to norm referencing more often than with conventional awards. Time constraints dictate that grading many (possibly up to 20) modules in a day is going to be far more difficult than grading 2 components, even though the individual module examination may be a lot shorter in duration than a component.

The simplest procedure, and one which was followed at the earliest grading sessions because of a paucity of data, is that (within the limits of the principal examiner's recommendation) each module should have a similar grade distribution and that this should be based on the experience of previous sessions. The problem with this is that the population of each module is not the same (and this will be explored in more detail under 'population') and there should be **no expectation** of equivalent distributions. It is debateable whether the distribution of the "hard" modules is enhanced to bring it into line with those that are easier or vice versa, but with this procedure little credit is given in the

awarding process for differential difficulty. The effect is hard to predict, but the suggestion is that it probably boosts the numbers gaining high grades on the earlier modules where the candidate population is very heterogeneous. With conventional assessment schemes the reverse is true, not only is the pattern of entry identical for each component, but usually there is also a high correlation in performances between components, especially for mathematics. There is thus far more justification for expecting the same, or similar, grade distribution for all components.

Rather more to the point is whether the variability of differential module difficulty is legitimate or illegitimate when the point is one of comparability. Clearly it depends critically on whether it is allowed for in the grading process: this is part of the investigation of later chapters. The starting point is, however, one of judgemental comparability.

Population

Superficially at least the candidates sitting an A level examination have much in common. However here there is a fundamental difference: the population sitting one component is, illness and accident permitting, identical to that taking all other components - they are also usually much of an age and have just completed a two year course of study. The population sitting any given module is, in general, not the same as for any other given module, although the extent of the heterogeneity will vary with the design of the modular scheme. Moreover, although there will be subsets of candidates for certification who have taken the same combination of modules, both in time and subject area, the "whole subject" candidates will often have taken a mixture of different (allowed) options. The effect on the grading of such a mixed population is complex and difficult to predict, but it must exist and as such would be deemed an illegitimate variability if its effects were not fully reflected in grade distributions.

An illustration of the problem is taken from the summer 1994 MEI Structured Mathematics syllabus for whom 4712 requested final certification i.e. entered the A-level. Of the six modules comprising this scheme, three are compulsory. This summer 8645 candidates entered module Pure Mathematics 1 (PM1). For only 597, just under 7%, of these candidates, was PM1 part of the terminal examination. In contrast, the entry for Pure Mathematics 3 (PM3), the third compulsory module with a similar candidature of

4468, included 2779, or just over 62% who were asking for final certification. For some of the later optional modules, 100% of the candidature would be final session A level candidates for whom the module would form part of the summative assessment.

The problem is compounded by the numbers taking both mathematics and further mathematics (and of course the single variants of statistics, pure mathematics or mechanics), for whom some modules could qualify for either examination. Since it is a truism that those sitting the "further" option, of whom there were 499 in 1994, are amongst the best mathematicians, their influence should be taken into account. It has been argued (Studies in Advanced Level Mathematics, 1988) that for awarding purposes those taking the double subject should be taken out of the mark distributions which are used for setting grade thresholds. How much more difficult to remove, say, all those for whom the sitting was not the final one for that module, even if it could be known. Again it would seem that such heterogeneity may work in favour of the weaker candidate, but to put much reliance on norm-referenced grading when the composition of the population is largely unknown could give rise to an illegitimate source of variability.

Maturity

The two complications which exacerbate any differences in the population have been alluded to above: these are the twin effects of maturity and resits, both legitimate variabilities. It is expected and encouraged that many, if not most, students will start taking modules in their first A level year (assuming the usual two year course). Some may even start in year 11 especially if they have taken GCSE in that subject a year early. They will not have the full range of A-level knowledge to draw on and experience of A-level questions will be limited. On average they cannot be expected to perform as well at these early sittings as they would later on in the course.

Resits

Not only are there these early takers with little experience, there are those who are re-sitting, worldly wise in the way of module testing, with experience not only of the examination itself, but being further along in the course also have an added maturity. Within the typical A-level course lifetime, the resit option is ostensibly dipolar - it is considered an inducement to weaker candidates, there is no "one and only" chance, but

a disincentive to stronger candidates for whom resitting to "get a better A" is not a powerful motivator and the stigma of failure, however slight, that the idea of re-taking an examination brings can be counter-productive.

However, the resit has been advanced as one reason why candidates achieve higher grades in modular examinations, if they do. Even though many schools only teach a module once, as in a conventional scheme, it is said that repeated assessments of the same skills and subject matter lead to a "greater command of substantive knowledge" (q.v. Ebel). Use of techniques learnt earlier during a course, and this could include vocabulary in the case of a language, familiarity with the examination (and some claim that a change of principal examiner changes the nature of an examination) will undoubtedly help induce such a greater understanding. Essentially it is a maturity factor, but a maturity factor which, unlike that of conventional schemes, brings with it A level examination experience. Notwithstanding the lack of both knowledge and skills which repeated sittings would imply, this better understanding and re-enforcement of knowledge by constant use should surely lead to a deservedly higher grade. Often, too, resits take place in a winter examination session with little pressure from other examination commitments when effort can be concentrated on a given subject area. It is noticeable that when early modules are resat terminally, they sometimes produce results lower than previously obtained (this is generally not the case over all). One explanation for this must be a lack of focus, especially when a reasonable result has already been obtained.

The one subject area of mathematics which has had modular schemes long enough to provide substantive evidence, and which is generally considered reliable in terms of marking standards certainly results in suggestions that there are improvements to be gained with re-sitting, moreover improvements which are not just the result of measurement error. Preliminary results suggest it is of the order of three UMS marks on the first resit and seven over two resits i.e. the more resits that are taken, the greater the improvement.

The resit is also one method by which a candidate's progress can be monitored if it forms part of the formative assessment. This ipsative aspect of the modular schemes is undoubtedly used by some centres who use it to determine future policy. There are also those who enter candidates for early modules with the aim of allowing them to gain

experience of the modular process, always with the idea of re-sitting at the next session. (This is clear from the results of some centres who, year on year, have very good subject A level results, but enter candidates for single modules with less than satisfactory results).

Of course, there are re-takers and immature candidates entering conventional examinations. The pattern of entry vis-a-vis these types of candidate is more or less constant year-on-year and certainly between components. The pattern of entry for modular examinations is different for each module, very few candidates for certification will be taking the first module as part of the terminal examination (at least for the first time, although there are still a few candidates who sit all their modules terminally), whereas there will be almost no re-takers for the more advanced modules. The SCAA rules now insist that at least 30% of a modular examination must be taken at the end of the course which effectively ends a practice whereby some of the better candidates had successfully completed all their modules before the final summer.

Opinions are divided as to whether resits should be allowed since they are perceived by some as unfair - or illegitimate. The position taken by the author is that if modular syllabuses are to be successful in enhancing the educational experience of A level candidates, then the resit is a legitimate and integral part of that assessment regime. The facility allows a more relaxed approach to the examinations, and provides an excellent diagnostic tool with which to improve future performances.

Length and Number of Examinations

Time and again (e.g. Nickson, 1994) the motivating effects of modular schemes have been cited as the reason for improved performances. There is a general acceptance that formal examinations in the first year of the sixth form tend to counteract the lethargy which occasionally sets in immediately after the effort expended in the pursuit of GCSE qualifications. There is little doubt that an impending module test in the first year of an A level course does, like hanging but not so terminally, concentrate the mind. However, such a test-bound curriculum is not always regarded with favour, especially by teachers.

It has been argued that questions at the beginning of an examination (ignoring the effects of differential demand) are the ones which produce the best performances. This

is because concentration is at its height, and tiredness has not become a factor. Many modular examinations, though greater in number, are shorter in duration than the usual 2½ or 3 hours, because of the limited module domain of behaviour. Thus, it is argued, a high level of concentration can be sustained throughout the examination leading to better performances and the cumulative effect over all modules may be manifest in an enhanced subject grade. However, one aspect of modular examining is that two modules may be time-tabled in a single morning or afternoon session, with just a short break in between. This would be equivalent in length to a traditional component examination with fatigue an equivalent factor.

Although the total number of examinations is often greater than those set for conventional examinations, the burden is generally seen as lighter than for conventional schemes. Some would also argue that the focusing effect of a limited syllabus examination, such as found in modular testing, makes it inherently easier. Although it would be possible to set up experiments to test such effects, they are not quantifiable within the confines of this research. It might be postulated that the length of the examination per se was a source of illegitimate variability, and, as such, would be considered by the awarders. Since grades are, however, based on evidence of attainment, it would be unfair to deny high grades simply because that evidence resulted from a shorter examination. Similarly, the limited content within the shorter examination would deny some breadth of evidence, but that might accrue from the totality of the examinations within a modular scheme, and grades must be based on the evidence of attainment from the given questions. These are matters of judgement, and the hypothesis is that judges are qualified to set standards which take these variabilities into account. One possible interpretation of the effect of the factors identified above would lead to the expectation of lower linear percentage scores for a given grade than would be required for modular candidates.

Question Structure

Questions may be of many types, for example multiple response, open ended, short answer structure and the demand of a question is dependent, in part, on the question type. In the scheme under consideration here i.e. mathematics, it is the declared philosophy within the modular scheme that all questions should be structured such that parts of each should be accessible to all candidates, though the whole only to a few.

This is clearly easier to sustain with a shorter examination, fewer questions and limited content, but if successful, would allow even the weakest candidates to achieve some marks. It is all too common in conventional examining that some questions cannot even be attempted by the less able, a policy which is clearly counter-productive and against the declared ethos of rewarding positive achievement, which, if it cannot be demonstrated, cannot be rewarded. There is also the counter to this argument that continual practice of questions found in linear schemes makes them easier. There are fewer 'compilers' and it is therefore easier to get to know a particular style of question.

Whilst the difference in question structure may be particular to mathematics, differences in question type are found between modular and linear schemes and are very difficult to quantify. Since it is often a matter of judgement as to which type of question is more demanding, it is incumbent upon the awarders to take differences in questions into account when setting grade boundaries. Because of the necessity of linking demand to evidence of attainment when determining standards, differences in question type, if not reconciled, would be considered a source of illegitimate variability.

Number of Examiners

Accepting the assertion by French et al (1987) that the quality of a candidate's performance is "a mental construct of an observer of that performance", is but a short step from accepting that the examiner's perception is in turn affected by the quality of the candidate's (s)he sees. The situation is clearly unresolvable - until the candidate's work is seen by an examiner, his/her quality is not known and a standard cannot be set, but that standard is a product of the quality of performance. The act of evaluating an examination script affects the standard by which it is judged. This is a phenomenon is well known in examination circles and has been investigated by Good and Cresswell (1988) when considering differentiated papers. The tendency of awarders to be lenient when setting grade boundaries on easier papers finds its equivalent in individual assistant examiners who tend to be more lenient in their marking if the preponderance of their allocation is of weaker candidates. In modular schemes there are many more examiners involved in the evaluation of the candidates' marks and therefore potentially far more perceptions of standards are aggregated to produce the whole.

It is not clear whether these uncertainties are to be considered as errors in the establishment of the true score on which all psychometric measurement is based, or whether there are to be a number of true scores. It is conceptually reasonable that every true score could have a reliability associated with it. Obtaining a value for such a reliability, which would be founded in some notional quantitative measure of the amount of ability being measured by that module, has always proved somewhat intractable.

The reductionist approach to modular examining forces consideration of the plurality of true scores. This is probably a rather different view from either the statistical or Platonic notions of true score (explained clearly in Rust and Golombok, 1989), but is nearer to the Platonic ideal than the statistical. However the statistical definition of true score leads to the general conclusion that the more assessments, the smaller the total error, and hence the "truer" the score. A modular scheme which usually allows many more assessments than those of a traditional examination would therefore give a "truer" measure of ability. But there is an additional complication consequent upon the partial-summative nature of the modular examination. The number of examination sessions at which a score can be delivered adds a temporal dimension to the notional true score which in a modular scheme becomes a function of time, and possibly resit number. However, it is not clear whether the same conclusion would follow if those measures were truly independent of each other.

Quite how the interaction between a module and an examiner may differ from that between a component and its examiner is not known. However, it is not only candidates who tire towards the end of a long examination, examiner fatigue can be just as influential on the outcome. There is no doubt it is easier to mark shorter module examinations - mark schemes are more easily learnt - and undoubtedly fewer errors are made both in marking and mark addition (a stupid but continuing source of error). If these factors point to a more generous examiner spirit, it is one that is unlikely to be proved.

In general then, the increase in examiners which automatically follows the higher number of assessment opportunities, favours no one candidate, but could lead to enhanced reliability, especially when one considers that many modules may include elements of both coursework and examination. It should not, therefore, be included in the variabilities to be considered.

Performance Profiling and Feedback

A clear source of a legitimate variability is the effect of feedback. It is incumbent upon examination boards who run modular schemes that all module results should be reported back to the candidate. It is the first time that such feedback has been allowed, and allied to the other formative assessments, provides information about performance which can, if used wisely, help with the selection of later modules. It would be difficult to devise an experiment to quantify the effect of such feedback; possibly a comparison of module forecast grades for candidates who were taking modules for the first time compared with ones which were resat with common first sitting forecasts as a control: but the number of resits testifies to its use.

Domain of Measurement

An illegitimate source of variability, this domain brings with it the considerations both of marks and grades. The processes by which the latter are determined are very different for the two schemes. The differences in setting the grade boundaries for component and module have been alluded to above, but that is only the beginning of the process which ends in candidates being given a syllabus grade.

Aggregation

Mechanistic rules often form a source of illegitimate variability, after the grade boundaries have been set. When components are graded, the final subject boundary for each judgemental grade is calculated according to the lower of two indicators (see GCE A and AS Code of Practice 1994 and Appendix A). The first indicator is simply the weighted aggregate of the component marks. The second is the weighted aggregate of the percentages reaching each of the component boundary marks. At the top end of the grade range it is this latter indicator which gives the lower mark because it allows for "regression to the mean", i.e. it allows a candidate with less than a minimum profile to obtain a given grade (e.g. a candidate with an A B B or even B B B profile can still get an overall A grade). However, in mathematics especially, the correlation between components is often high and in this case there will be little difference between the two indicators and regression to the mean effects small. In one or two subjects e.g. geography, where the correlation between components is low, the difference between

the two indicators can be as much as 4% of the total. Additionally, once component boundary decisions are known, they can be mapped in their two forms - mark aggregate and percentage aggregate - on to a total mark distribution for the final syllabus boundaries to be calculated (see Appendix A) and the effect of grading decisions known immediately.

Every modular scheme requires a two stage aggregation process - from raw marks to points/uniform marks/credits etc before a final syllabus total can be calculated (a fuller description of this process is given later in the next chapter on the syllabus to be investigated) and will almost always involve marks obtained in other examination sessions. Thus pragmatically, if for no other reason, each module must be considered as a stand-alone element within the regime and no reference should be made to the grade distributions of other modules because of the differences in population. (It is possible for some examination processing systems to produce module grade distributions for common candidates on which comparisons can be made, but not currently in OCSEB). It is not practically possible to attempt any form of aggregated percentage, and no allowance for "regression to the mean", in the strictest sense, can be made by this method. SCAA rules allow a shelf life for modules for up to four years and any attempt to aggregate percentages with modules spread over a four year period is bound to be fraught with problems. Therefore the grading decisions which awarders must make are fundamentally different from those made for non-modular schemes.

The distinction between the two processes is not always understood. Some modular schemes allow for regression in the aggregation (as was the recommendation from Thomson's report) simply by adding the uniform marks or points and allowing the boundary to be less than the minimum module score (for the grade). For example UCLES uniform marks system for A levels was based on a 6 module system. Each module uniform marks were out of 100 such that the boundary for each module was given by:

A	B	C	D	E	N
80	70	60	50	40	30

but the syllabus grade boundaries were

A	B	C	D	E	N
450	390	330	270	210	150

such that there was an offset or allowance of 30 marks at each boundary.

So to phrase the difference merely as a matter of "allowing for regression" in the awarding decision i.e. pitching the boundary slightly lower than one would for a comparable exam were it a component, is to miss the fundamental dichotomy. Awarders on non-modular schemes expect similar grade frequency distributions for each component. There is a tendency for awarders on modular schemes to expect the same without the same underlying justification - namely the same candidature and often the same subject matter. This latter is a consequence of the modular approach, and most modular methods of assessment include many more modules than their orthodox counterparts include components.

Regression and Compensation

Regression, or rather an allowance for regression, and compensation are two areas where disparities can arise between linear and modular schemes, even when all other aspects of the assessment are the same. Compensation is the ability of a good score to counteract a worse score. Because modules may be taken over a period of time, and it is necessary to use a conversion to standardised scores in order to aggregate modules fairly, the conversion can affect the ability of a good score to compensate a bad one. This is because the 'value' of a raw mark changes with conversion, and that change depends critically on the grade on that module. Also, the bunching effect of aggregation, (described more fully in appendix A) which is counteracted by using percentage rather than mark aggregation at high grades in linear schemes is not available to modular awarders. They must make 'implicit allowance' for regression when boundaries are determined.

It is argued that any system of aggregation for modular schemes should retain the power of a high mark to retain its compensatory power. The rationale for this allowance may be somewhat equivocal, but the only sensible way it can be accommodated in a modular scheme is at the awarding stage. The more modules there are, the more potential for compensation and it is a balancing act between the setting of grade

boundaries and the aggregation process to get a suitable final mark/point for each module.

The other aspect of determining a suitable aggregation process is the effective raw mark scale that is actually used. In English, for example, it is rare for any examiner to use the top end of the mark scale, and almost unheard of for any candidate, however outstanding, to be awarded full marks. The same is not true of, say, mathematics. Thus there is an argument which promulgates the concept of mapping the raw mark scale that is actually used onto the uniform mark or point scale. This would allow the compensatory power of the "best" score to be the same, whatever the subject.

The cumulative effect of the different grading policies is unpredictable, but the compensatory effect of the higher scoring modules, usually those taken more than once, should not be underestimated. It is, however, entirely possible that if modules are considered as simply components under a different name, loss of both effective compensation and regression allowance, the number of high grades will decrease at syllabus level. Whereas in linear schemes it is possible to accept very different grade distributions across components because one can be adjusted so that the final result is as required for year-on-year comparability, the essential balance necessary for modular awarding is often not fully appreciated or understood. A full discussion of these issues will be found in appendix B.

Generalisability

One final issue to be considered is how applicable may any results be to other syllabuses. The methodology and assumptions are viable ways of investigating the comparability of other subjects. However, mathematics, the subject investigated here, has particular characteristics which not only make it a suitable vehicle for modular assessment, but also may inhibit the generalisability of any findings. Prime amongst these is the 'linear' nature of the subject. Put crudely it gets harder the more progress that is made. This means that the early modules are easier. In other subjects where skills such as language are very dependent upon maturity, early modules may lead to poor results. In mathematics the later modules tend to attract lower grades, one would not expect the same in English, say. Therefore the pattern of behaviour is likely to be different for different subjects.

However, that does not mean that all the findings are invalid for other subjects. If grading standards can be maintained, then the variabilities may be qualitatively the same, though quantitatively different. Mathematics is one of the subjects least susceptible to measurement error, English one of the most, and the effects of resits could well be very different. Also because of the high correlations between components the allowance for regression in mathematics is very small (here, zero) and mechanistic differences are fairly minimal. UMS conversions are designed with a view to the expected outcomes and although consistent within the scheme, may not translate to other subjects without distortion.

Discussion

The preceding discussion points out those areas where major differences lie, and also indicates where the nature of the modular scheme could, quite legitimately lead to different performances and possibly better than expected grades. Even if only a few marks are gained from each of the factors, cumulatively the consequences could be quite noticeable. In the early days of modular schemes it was felt that these factors would offset the regression allowance, but, of course, if candidates have not used the resit facility and are fully mature and even choose to sit all their modules terminally then they could be positively disadvantaged by the modular aggregation process.

The defining characteristic of a modular scheme lies in the separate domains of behaviour and assessment ascribed to each module. This affects the demand of the examination in ways which are difficult to quantify. Since demand and attainment are inextricably linked, the processes by which a modular grade is achieved are different from those of linear schemes, as are the constituent parts which contribute to the make-up of that grade. The commonalities of a common core and final grade are not, by themselves, enough to beget comparability, nor are the processes by which the grades are determined. But because the grades can be put to the same use, e.g. qualification for university entrance, they are endowed, by the end user, with a social equivalence which is the central theme of this thesis. That the grades are arrived at by different processes does **not** invalidate the methods by which those grades are achieved. It may be thought that the different approaches are examples of different syllabus genera, with their own, separate environments inimical to each other. That the outcomes are

considered comparable is almost inevitable because of the award of a grade, and it is this commonality which invites investigation of the processes by which they are achieved.

Thus the question which is usually asked is the wrong question. It is not "should modular schemes of assessment be comparable with traditional methods?", because in at least one sense, the social, this can be assumed, but "are modular schemes valid means of assessing attainment at A-level?". It is precisely because they have, in practice, been considered comparable and valid (see the Cresswell social definition of comparability), despite reservations on the part of some educationists, that there needs to be an investigation into the construct irrelevant variances, here defined under the heading of variabilities, and construct under-representativeness of the schemes. Making the assumption that standards are the same within and across schemes of assessment, allows differences, where they are found, to be quantified and labelled. If, by taking legitimate variabilities into account, differences in achievement by equivalent candidates can be adequately explained, then the assumption of comparable standards must hold. If, however, those legitimate variabilities can only go some way to explain discrepancies in expected outcome under two schemes of assessment then it is the illegitimate variabilities which have produced an imbalance in grading standards, i.e. a failure to take into account all those differences in any two sets of assessment when reviewing the evidence of attainment which determines the grade standards.

In summary, the key proposition of this thesis is that by investigating legitimate and illegitimate variabilities as defined in this chapter, it is possible to ratify, or not, the underlying assumption of comparable grading standards both within and across linear and modular schemes of assessment in the same subject. The remainder of the study is therefore concerned with the investigation of legitimate variabilities, and, inter alia, some of the more accessible of the illegitimate differences, in an attempt to quantify such differences as are seen. Finally, by using a prior measure of achievement, it is possible to quantify any advantage which appears to accrue from following one particular type of assessment scheme. A crude measure of the variability due to illegitimate sources which might lead to an undermining of the original assumption of comparability can then be found by comparing expected differences from the two sources.

CHAPTER 5

A Perfect and Absolute Blank - Within Subject Comparability

There are potentially more variabilities that enter modular assessment regimes than linear ones because of the number of examination opportunities which are open to the candidates. This chapter investigates how the flexibility of one modular scheme affects behaviour and how the variabilities therein may influence final grades. The base position is that of comparable grading standards across and within modules, and one of the main considerations is the use of resits and how they affect scores.

Modular schemes have been in development for a number of years and a number of different assessment structures have been suggested (SRAC, 1990). Of the two which have been developed, only one is distinct (that defined in Chapter 4). The quasi-modular or hybrid schemes are a half-way house between linear and true modular schemes with the assessments often taking place terminally. For some time it was suggested that schemes could be both linear and modular depending on when the examinations were taken. It was a suggestion which did not last long and only two distinct types of assessment now exist at A level, linear or modular.

Of the modular A levels which currently exist, few have reached an "ideal state", i.e. they have not reached that stage in their development when they have a number of centres who have been entering candidates for several years as well as continuing to attract an increasing entry and no fixed patterns of behaviour have been established. Of those which could claim to be in such a position, only one could also claim the flexibility of module choice and use of the type of aggregation model which may be used for all modular syllabuses in the foreseeable future. That syllabus is the Structured Mathematics Syllabus devised by MEI and administered by the Oxford and Cambridge Schools Examination Board.

The reasons for this choice are simple:

(i) it is the longest standing of the OCSEB modular schemes and attracts a large enough entry for detailed investigation

(ii) it uses the same aggregation method as other schemes, albeit to a slightly different scale

(iii) it was designed as a modular scheme from its inception and embodies those principles of modularity which distinguish it from a hybrid scheme.

It is a leader in the field of modular examining, though it was neither the first nor the most innovative. However it has provided the prototype for examinations which are to follow, not only for the Oxford and Cambridge Board, but also for all other Boards who have instituted modular exams of their own. Because of the high entry it is possible to investigate effects which are too small to be seen in other regimes. The first certification for this syllabus was in 1992, but such is its appeal that from a candidature of 763 from 52 centres in 1992, it has grown nearly six-fold to 4371 from 219 centres.

The 1994 syllabus on which this thesis is based requires the results from six modules, three of which are compulsory, to be aggregated to form a syllabus total. Any optional module may be taken provided the rules of dependency are adhered to. (These are given with a schematic representation in the 1992 MEI syllabus.) Although this thesis is only concerned with the A level entitled "Mathematics" and which thus includes the subject core, it must be noted that a different arrangement of modules could produce certifications of Pure Mathematics, Mechanics or Statistics, without even considering the Further Mathematics variants.

Up to and including 1994 all terminally assessed modules carried a maximum mark of 60, which for some was split between 10 or 20 coursework marks and 50 or 40 written paper marks. They can all be resat within the lifetime of the course. Each module has its own assessment scheme, usually including a one hour examination which may allow some question choice. There are 22 separate modules, four of which are wholly assessed by internal assessment or coursework (these carried 100 marks each with fixed boundaries), but there have only ever been entries to 20 of them.

Aggregation is by uniform marks. This system is also known as a standardised score or fitting marks to a common spline. Module grade boundaries are determined by the same method as component grades using the raw mark scale. This is then mapped into a uniform mark according to the following rules. All grades except A (11) and U (9) have a spread of 10 uniform marks, such that the A boundary is 60, B 50 down to U with a maximum possible mark of 70. Thus the uniform mark thresholds are

A	B	C	D	E	N
60	50	40	30	20	10

These remain unchanged whatever the raw mark boundary. The conversion is achieved by mapping the raw score onto the correct uniform score. For example suppose the raw grade boundaries for a certain module are (introducing the convention that small letters are used for module scores)

a	b	c	d	e	n	u
43	35	29	23	18	11	0

and further suppose a candidate had gained a mark of 25 on this module. This is clearly a 'd', but in order to aggregate it with other module, the mark must also show how good a 'd' it is. The range of marks in the d band is 6, so 25 is 2/6 of the way towards a 'c'. Mapped onto a range of 10 for the uniform mark, the final UMS (uniform mark score) will be given by the D threshold (30) + 2/6 of the D mark range (10), i.e.

$$30 + 2/6*10 = 33$$

It must be emphasised that there is no allowance for "regression" and the compensating power of an A grade over all other grades is small (1 mark). The final grade is calculated simply by adding the six best module marks with the UMS syllabus boundaries merely set at six times the individual marks:

A	B	C	D	E	N	U
360	300	240	180	120	60	0

This is felt to be appropriate for mathematics where maximum marks are achievable and are achieved. It should also be recognised that the UMS is the same order of magnitude as the raw mark scale. It alleviates any problems associated with "premature approximation", that is allowing the raw mark scale to become so compressed on conversion that too much information is lost and erroneous grades can result (q.v. appendix B). Additionally, MEI papers have always been set with the idea of design thresholds. These are related to the UMS scores and, provided the evidence of attainment at these thresholds is very close to the standard required for a grade

boundary i.e. the chosen boundary is approximately equal to the design threshold, conversion rates will be kept as the designers intended. They will not vary much from session to session and conversion rates will be fairly constant. Problems do arise when conversion rates for different grades vary, and it does highlight the need for question paper setters to take into account the effect of UMS conversions by designing papers so that expected thresholds are close (in percentage terms) to those of the UMS scale.
The focus of both this and the following chapter is to describe the expected behaviour of a typical A level examinee of the future (when many more, if not most A levels will be modular) using data from the 1994 structured mathematics scheme as a model (details of which will be found in the MEI Structured Mathematics Syllabus, October 1992),

Another, equally important, aspect of the analysis is that of between module comparability. Whilst it is true that strenuous efforts are made to ensure comparability between components in linear examinations, the need to ensure year-on-year comparability being paramount, it is not unknown for adjustments to be made at component level in order to achieve this, even at the expense of strict component equivalence. It is an adjustment not available to modular awarders for two reasons: firstly, they may well not know the syllabus outcomes of their decisions until after the award meeting and secondly, since the aggregation process will include modules awarded on earlier occasions, it is imperative that module comparability is maintained otherwise candidates reaching different module standards could gain the same final syllabus mark. This would be palpably unfair.

Another consideration is module choice. In most, if not all, modular A levels there is a degree of choice allied to some compulsory elements. Arguably, it is only strictly necessary to ensure temporal equivalence for the compulsory modules with comparability both across time and module for those which are optional. However, in the interests of balance, it is preferable that there be strict module comparability between all modules. One indicator of this is from module pair results, but there are ramifications not attendant in conventional schemes. Enhanced module performance may come from a resit, or more maturity, and these need to be considered as separate entities before module pair analysis can be interpreted with rigour.

Therefore, in an attempt to build up a picture of the behaviour of a modular candidate, the influence of uniquely modular complexities must also be taken into account. There is

an additional difficulty which is only found in mathematics examinations. That is the presence of two cohorts within one examination - i.e. single mathematicians and double mathematicians. This is particularly problematic when trying to build up a picture of module choice since those asking for certification for further mathematics will be taking double the number of modules. For much of the analysis therefore these candidates have been excluded. In the past, there has also been the suggestion that for awarding purposes, further mathematicians should be excluded from the mark distributions i.e. statistical indicators should be derived from performances of single mathematicians only on the grounds that the inclusion of further candidates at the top of the mark range unfairly depresses the number of high grades awarded to the single mathematicians. It is equally a complication for modular awarders.

Some Basic Data

In 1994 there were some 21 modules which could contribute to an award within the MEI Structured Mathematics scheme. Of these 21, six are set on Pure Mathematics subjects, PM1 to PM6, six on Mechanics M1 to M6 and six on Statistics S1 to S6. The final three comprise Decision and Discrete Mathematics (D&D), module 19, Commercial and Industrial Statistics (C&IS), module 20 and Numerical Analysis (NA) module 21. Although there are a very few candidates who take up to 18 modules (for the award of three A levels), the concentration of this chapter is on the single award entitled "Mathematics", for which six modules, three compulsory, need to be completed. The published data for 1994 show:

Table 5.1: 1994 Grade Distribution for Modular Mathematics

Grade	Number	Percentage	Cumulative Percentage
A	1450	33.3	33.3
B	1087	24.9	58.2
C	865	19.8	78.0
D	577	13.2	91.3
E	338	7.8	99.0
N	42	1.0	100.0
U	1	0.0	100.0

By way of contrast, the OCSEB traditional mathematics syllabus had the following grade distribution for 1994:

Table 5.2: 1994 Grade Distribution for Linear Mathematics

Grade	Number	Percentage	Cumulative Percentage
A	338	37.1	37.1
B	181	19.8	56.9
C	136	14.9	71.8
D	110	12.1	83.9
E	56	6.1	90.0
N	49	5.4	95.4
U	42	4.6	100.0

Apart from the obvious difference in cohort sizes, which for the modular scheme is on the increase though decreasing for the traditional scheme, also evident are the higher B and C rates of the modular scheme. Even allowing for differences in candidature (which traditionally is mainly from the independent sector though increasingly not so for the modular scheme) it would seem at first glance that more candidates than expected were gaining A to C grades from the modular scheme.

Another, slightly negative point to note about these figures is that there are so few U grades awarded in the modular scheme. In most subjects there are usually more than this and in other mathematics syllabuses the proportion of U grades is 4%-5% as shown above. These figures disguise and also highlight one of the more intriguing aspects of modular examinations. Because the element of choice plays such a large part in this type of scheme, some candidates choose to exercise the ultimate, and unique right of withdrawing from certification.

In fact the total number of entries for this modular syllabus was 4671 candidates, of whom 6.7% were designated as X, a category of candidates for whom no grade (including U) was awarded. Within this category there would be some absentees, but there were more pertinent reasons for inclusion. Prominent amongst these were too low a final grade, i.e. an "underperformance" in the terminal examinations which resulted in

a lower than expected award. Thus verisimilitude is added to the oft quoted reason for "too many high grades" being awarded to modular examinees, namely the withdrawal option being exercised. For example, inclusion of the X candidates in the total would reduce the percentage of As above by 2.5%. The conventional A level candidate does not have this choice.

There were, however, other, more bizarre reasons for an X designation the most common being a mis-entry either because the school had got it wrong (e.g. one centre's candidates were entered for A level when they had only done three modules each) or the candidate had not sat the correct combination of modules in which case their certification may be made under another title, e.g. "Statistics".

Not all Boards allow candidates to choose to withdraw from certification in this way, in which case the anomaly disappears; other Boards have a policy of "sweeping" the module data-base and may award certification on completion of the correct numbers of modules whether requested or not - this could lead to a jump in numbers "taking" AS examinations. (SCAA rules allow modules to be re-used in some circumstances, many AS awards made in this way are, in fact, purely nominal). It is difficult to defend either policy given the spirit of modularity and its intrinsic quality of choice.

Also included in the above figures are the candidates who also chose to take a further mathematics A or AS level. Although this option has been the subject of some dispute over the years as clearly a further mathematics candidate is expected to perform at a higher level than for the single subject mathematics and thus the further mathematics A level is usually considered to be of a "higher" standard. This is invariably borne out by their results for the single subject: here, for example, over 90% of the 516 candidates who went on to take further mathematics A level obtained an A for mathematics, with three quarters of the 227 AS candidates also getting A for the single subject.

Although these candidates were legitimately included in the overall grade distribution, they formed a sub-group which distorted the data. When, for example, calculating which are the most popular modules for those sitting mathematics, it must be remembered that the inclusion of those sitting further options adds to total module numbers. The "mix 'n match" method of calculating grades for more than one syllabus entry means that some modules are transferable between options and are often subject to manipulation to

produce the "least best" result; that is the lowest total that could be found (within the rules) which would give the best possible result. An example from the 1994 examination illustrates this point.

The candidate gained the following module profile:

Module	1	2	3	4	5	7	8	9	10
UMS	68	63	62	58	56	60	48	50	35

Module	13	14	15
UMS	50	52	46

Modules 1, 2 and 3 must be included in the single subject total, but, even allowing for the rules of precedence the other modules could be combined in several different ways. In the example above, modules 1, 2, 3, 4, 7 and 13 were combined to give a total of 361, an A, and the rest combined for a total of 287, a C grade. In a different legitimate combination (1, 2, 3, 13, 14 and 15 or the rest) this could also equal two B grades. It was often possible (though not with these scores) to combine them in such a way that would give for example a very good A, here the maximum A score would be 367, leaving such a low score for the second subject that a grade could be lost. The general rule is to award the highest possible grade, here A, for the lowest possible score, here 361.

Clearly none of these calculations impinged on the behaviour of the typical examinee for the single subject and for some of the analysis below all further mathematics A and AS candidates are excluded. The full following grade distribution is for single subject candidates only and includes the X candidates:

Table 5.3: Grade Distribution for Single Subject Modular Candidates

Grade	Number	Percentage	Cumulative Percentage
A	810	20.6	20.6
B	1010	25.7	46.3
C	849	21.6	67.9
D	576	14.7	82.6
E	338	8.6	91.2
N	42	1.1	92.3
U	1	0.0	92.3
X	302	7.7	100.0

Of these candidates about 50 failed to complete the full set of modules to gain an award. Of the rest, most chose to withdraw thus allowing them to resit the terminal modules only in an attempt to gain a higher grade overall. Some however failed to reach the hurdle of 5 UMS marks (i.e. min(N)/2) in each module which automatically put the candidate into the X category.

In as much as rules of combination and nomination of modules, re-use of modules (requested or otherwise) and post hoc withdrawal from certification vary between different schemes of assessment, to use MEI mathematics as a paradigm for all modular schemes could be questioned. However, because the average number of *different* modules sat by single subject candidates in 1994 was 6.1, indicating that most candidates only took the minimum number of modules, and the majority chose certification, it is unlikely that any of these variations would affect the more general conclusions.

Patterns of Behaviour within Modules

In order to build up a pattern for modular schemes it was necessary to investigate typical entry behaviours. It was tempting to include both X category candidates as well as those taking a further mathematics option in order to enhance the data. But in the former case the module pattern was likely to be incomplete, and in the latter too many modules

would be taken. In neither was the pattern typical of the single subject entry, and thus both types of candidate were excluded from the first part of this section of the analysis.

One fairly unusual feature attaching to mathematics schemes was that some candidates appear to take three years to complete the course. Of those candidates certificated in 1994, 8% started sitting modules in the summer of 1992, with 7 candidates the previous winter. The majority of these were candidates who sat their mathematics GCSE a year early and whose centres chose to embark on the modular syllabus study in year 11 instead of, say, additional mathematics O level, though there would be a few who were re-taking terminal modules from summer '93. Whilst it is conceivable that modular syllabuses in subjects such as modern foreign languages (MFL) which also attract a number of early entries (though usually only in French) could extend from year 11 to year 13 (inclusive), it is unlikely that it would be common in subjects such as science unless patterns of entry at GCSE changed considerably. The pattern of entry has been considered separately for each module.

Pure Mathematics Module 1 (PM1)

This module was one of the three compulsory elements, the others being PM2 and PM3, which together constituted the core syllabus. This, by inter-Board agreement, must be included in all A levels entitled "Mathematics". There were two examination opportunities annually in the winter and summer of each year when module examinations are set, but a syllabus grade could only be awarded at the summer session. SCAA rules which dictate that 30% of any A level assessment must be terminal effectively meant that no student could complete their modules at the winter session. In the summer '92 examinations only the first module of six in each discipline (pure mathematics PM, mechanics M or statistics S) attracted any entry of size from those graded in 1994, with the largest being PM1 at 7.6%. This figure was an order of magnitude below that for summer '93 when the largest proportion (78.5%) of those candidates sat PM1. Table 5.5 gives the numbers of resits for each module. Not surprisingly PM1 was one of only four modules which were resat as much as four times, although perhaps less predictably, the majority of candidates only took the module once, with only 40% choosing to resit. Most candidates took the module in the summer 93 session, leaving them a further year for resits. The explanation might lie at the extremes of the ability range; good candidates performed well enough to nullify the need for a

resit, whereas weaker candidates would not sit this module until later in the course, leaving them less time to resit.

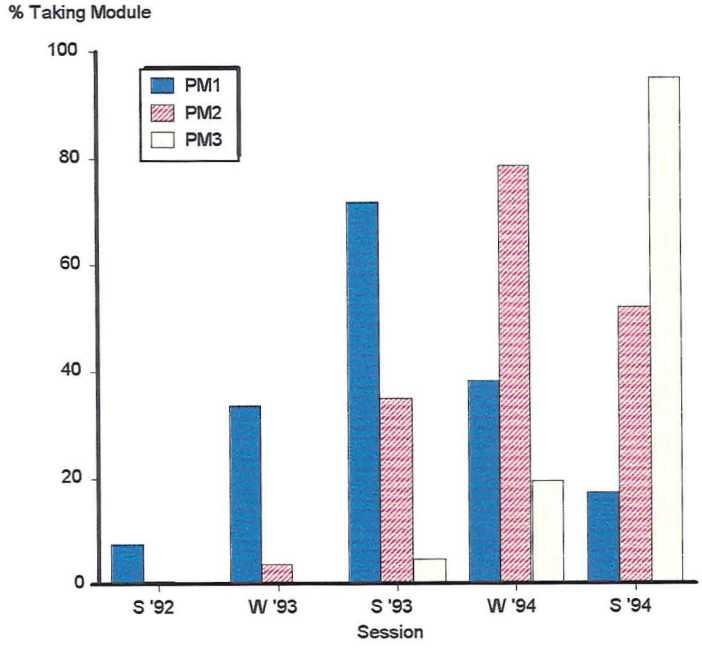
Pure Mathematics Module 2 (PM2)

The pattern of entries to this module might be predicted given the progressive nature of this part of the scheme. Reference to table 5.5 shows that about 6% more candidates resit this module, with the majority taking it in the winter before certification, although over 50% still take it terminally. The same entry pattern as PM1 over three sessions was observable, albeit displaced by one session.

Pure Mathematics Module 3 (PM3)

Almost all candidates (over 95%) sat this module terminally, although for about 10% of them it would not be their first attempt. Only 19% sat it the previous winter, and most did not resit. Figure 5.1 illustrates the entry pattern for the compulsory modules over all recorded sessions.

Figure 5.1: Proportion Taking PM1, 2 & 3 by Examination Session



It is worth recording here that the mean grade for candidates taking these modules (as well as the mean overall grade) is 5.35 (taking A as 7, B as 6 etc), i.e. the average candidate was most likely to have obtained a C grade.

Pure Mathematics Modules 4, 5, and 6 (PM4, PM5 & PM6)

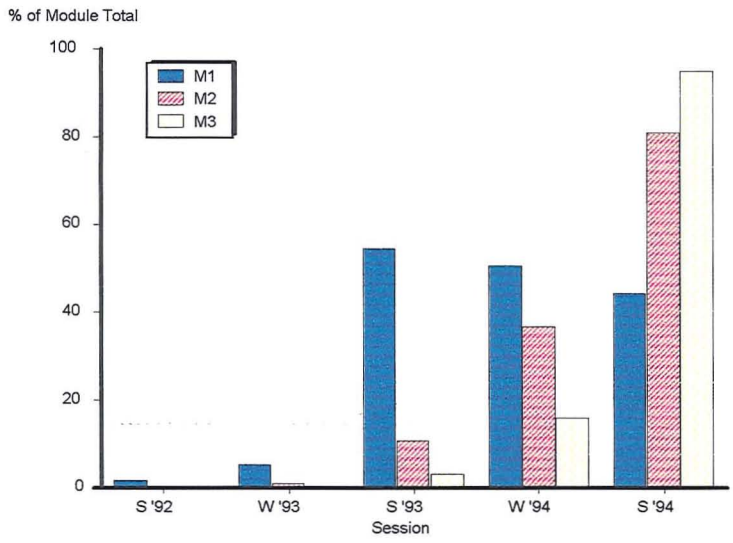
The sessions at which these modules were taken was of somewhat less interest than the number choosing them. Also, given the rules of dependency and the natural progression within the subject it was unlikely that any candidate would take any of these options before PM3. It was possible within the rules (though not recommended) to sit these modules independently of each other although it should be noted that a full suite of pure mathematics modules would lead to certification under the title "Pure Mathematics" not the mathematics syllabus on which our interest was focused. Almost all the candidates for these modules sat them once and terminally. Only 76 candidates took module PM4, for whom the average overall grade was 5.4, marginally above that for all candidates. PM5 was taken by only 7 candidates, average grade 6.6, and PM6 by a single candidate who gained an A. All these, but especially the latter two, modules clearly attract a very small entry from atypical candidates, and as such could be discounted when trying to construct a picture of an average modular examinee.

Mechanics Modules 1, 2, and 3 (M1, M2 and M3)

None of the "further" mechanics modules 4 to 6 were taken by any single subject candidate, although the first mechanics module was taken by 76.9% of all candidates, M2 by 45.3% and the third by 10.3%. The mean grade for M2 was slightly higher than for M1 or M3 implying that after taking the first mechanics module the slightly weaker candidates on average dropped mechanics, but those choosing to take the third module were slightly weaker than those who took only two. Reference to table 5.5 shows that the resit pattern for M1 was very similar to that for PM1, although for both M2 and M3 over 80% took the module only once, with no-one taking M3 more than twice. Figure 5.2 shows the pattern by entry by session for the mechanics modules. The percentages are taken from the total taking that module *not* as a proportion of the total candidature for single subject mathematics.

It can be seen from the figure that whereas the percentages taking M2 and M3 increased over the examination sessions, that for M1 reached its apogee in the Summer of 1993 session. There were obviously contrasts with the compulsory modules, but entirely predictably later mechanics modules attracted the greater proportion of their entries towards and at the end of the course.

Figure 5.2: Proportion Taking M1, 2 & 3 by Examination Session

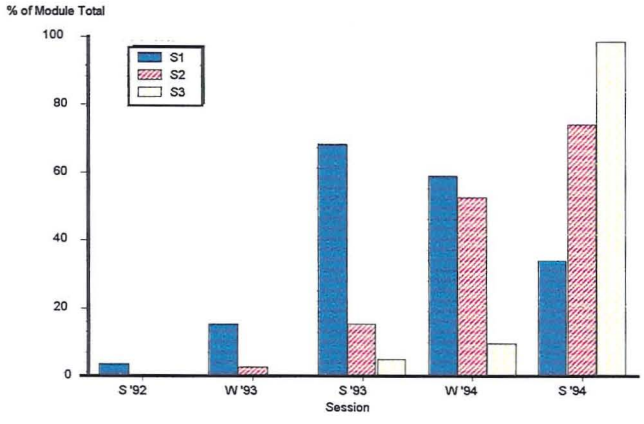


Statistics Modules 1, 2 and 3 (S1, S2 and S3)

Statistics 1 was the most popular of the optional modules and was taken by over 90% of the single subject candidates. The second statistics module was slightly more popular than its mechanics counterpart and was taken by over half the candidates. Statistics 3 was taken by twice as many candidates, 21%, than Mechanics 3. Given that, of all the optional modules, a higher proportion sat the S1 examination in the first summer of the course, it was not particularly surprising that statistics 1 was also resat by a higher percentage than the respective percentage for any other module. In fact this was the only module where more than 50% of the candidates chose to resit. Because nearly all the candidates take S1 it was not surprising that the mean overall grade for candidates sitting this module was 5.34, but it was marginally lower, at 5.3, for both S2 and S3. This

indicates that it was generally the weaker candidates who chose the later statistics modules. What was not clear was the effect of the module itself on the final grade, a severely marked module would depress the final grade. It remains to be seen from later analysis what the relative severities/leniencies of each module were.

Figure 5.3: Proportion Taking S1, 2 & 3 by Examination Session



The graphical picture which emerges shows a similar pattern of entry over the various examination sessions as was found for the three mechanics modules. Again given the nature of the subject, especially when sub-groups were considered it would be somewhat surprising if such similarities were not observed.

Decision and Discrete Mathematics and Numerical Analysis (Modules 19 and 21)

These two were the more unusual choices which were made. Both consisted entirely of coursework with no terminal examination. In 1994, 9.1% of the candidates chose to undertake the investigations posed by the decision module, achieving an average overall grade of 4.92. Clearly these were amongst the weaker candidates in the cohort, but again there was the caveat of cause and effect which remains to be investigated. The numbers completing the module at the various sessions started from less than 1% at the winter '93 session, and rose steadily at each session to 83.9 who submitted to judgement in the summer of 1994. Only 12 candidates took numerical analysis, all in the final session. However their average grade was 6 which was high compared with all

other modules. Only further investigation will show if this was entirely justified given the module results.

Real Differences

Using the Kolmogorov-Smirnov two-tailed test for ordinal data, with p <= 0.05, a table was constructed showing those modules whose candidates' syllabus grade distribution proved significantly different from each other.

Table 5.4: Modules with Significantly Different Syllabus Grades

MODULES								
	PM1-3	**M1**	**M2**	**M3**	**S1**	**S2**	**S3**	**19**
PM1-3		0	1	0	0	0	0	1
M1			1	0	0	0	0	1
M2				0	1	1	1	1
M3					0	0	0	1
S1						0	0	1
S2							0	1
S3								1
19								

Key: 0 indicates no significant difference found p <= 0.05

1 indicates that significant difference exists between the modules p <= 0.05.

This table shows clearly that the two modules whose candidates differ significantly from those for other modules are M2 and D & D, module 19. M2 candidates seem to get higher than average results, whereas D & D candidates are the weakest found for any module. How much influence the grading of the module itself has on these grade distributions has yet to be investigated.

Resits

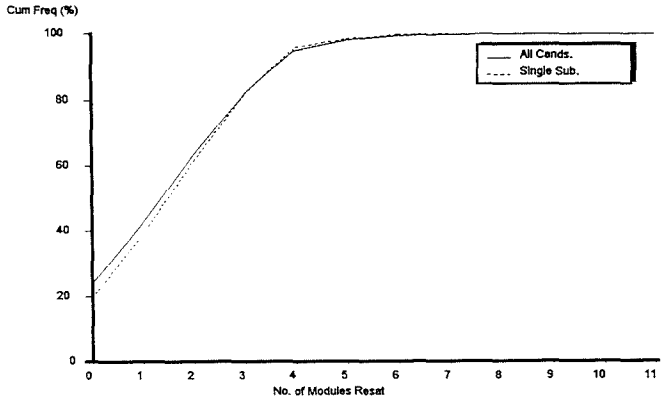
Whilst the interest in the foregoing section lies in the popularity of individual modules and the pattern of module taking which would be obscured by the inclusion of double

award candidates, it is more instructive to consider resit patterns as they exist for all candidates i.e. those taking either single or the double award. Although each module has been considered separately often with reference to the single subject data given in table 5.5, when presented together a cumulative picture emerges. There are essentially three questions which have to be addressed:

i. how many modules do candidates resit
ii. how often are individual modules resat
iii what mark gains are achieved by resitting modules.

Figure 5.4 illustrates the pattern of resitting by candidates. Since it is a cumulative frequency graph it illustrates that just under a quarter of candidates had no resits, about 40% resat either one or none (i.e. 20% resat once), about 60% resat none, one or two modules, with very few candidates resitting more than four modules.

Figure 5.4: Cumulative Percentage of Modules Resat



The graph shows the steep, but constant rise to about 95% of candidates who took 4 or fewer resits where it then levels off. Also indicated is that over 60% of candidates resit less than three of their modules. The superposition of the data for single subject candidates only shows that where resits are concerned there is little observable difference in the pattern of behaviours for the different types of candidate.

The second point to note, from table 5.5, is that each module, except S1 for single subject candidates, is taken only once by the majority of candidates. Also resits were

independent of ability as defined by the final grade, i.e. the pattern of resits was very similar for all grades (although grade distributions are not given here). Obviously those modules which naturally fell at the beginning of the course of study were more likely to be re-taken, and this was clear from the results.

Table 5.5, which has been referenced previously for single subject candidates, gives, as a percentage of the total for that module, those candidates resitting each module; e.g. 62.7% of those candidates taking module 1 (in fact the whole cohort) only sat it once, i.e. no resits; 26.5% had one resit, i.e. took module 1 twice, etc.(the results for single subject candidates only are given in brackets where they differ).

Table 5.5: Resit Pattern for Each Module

	Resit Percentages				
Module	**0**	**1**	**2**	**3**	**4**
1	62.7(59.9)	26.5(28.4)	8.8(9.5)	1.9(2.1)	0.1
2	58.1(53.8)	32.9(36.5)	8.2(8.7)	0.8(0.9)	0.0(0.1)
3	89.3(88.8)	11.3(10.3)	1.5(0.9)	0.0	-*
4	86.7(93.4)	12.5(5.3)	0.8(1.3)	-	-
5	91.5(85.7)	7.9(14.3)	-	-	-
6	100.0	-	-	-	-
7	68.0(64.7)	24.4(27.1)	7.1(7.7)	0.5	-
8	80.1	16.8(17.4)	2.8(2.3)	0.2	-
9	91.8(89.8)	7.5(10.2)	0.6(0.0)	-	-
13	51.9(48.4)	35.7(38.3)	11.7(12.5)	0.7(0.8)	0.0
14	72.1(69.7)	24.9(27.2)	2.9(3.0)	0.1(0.0)	0.0
15	93.4(94.5)	5.5(4.7)	0.9(0.8)	-	-
19	96.1(96.0)	3.9(4.0)	-	-	-
21	100.0	-	-	-	-

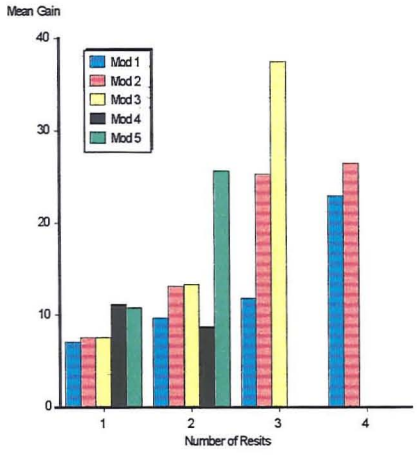
* Denotes no candidates in this category.

The table also bears out the assertion that although there is a greater tendency for single subject candidates to resit, the difference from the whole cohort figures is small

enough to indicate a similarity in resit behaviour between single and double award candidates.

The figures below illustrate the average mark gains on resit for all candidates. Whilst it would have been possible to show the average maximum mark difference, this would disguise the pattern of mark gain (which can be negative if there is a loss) determined by the number of times the module has been taken. Therefore, if, for example, a candidate has resat a module 3 times, then their data are also included three times, once for the gain on the first resit, for the total gain on the second resit (i.e. the gain over the first time the module was taken) and finally for the mark difference between the third resit and the initial result.

Figure 5.5: Gains on Resit for Modules 1 to 5

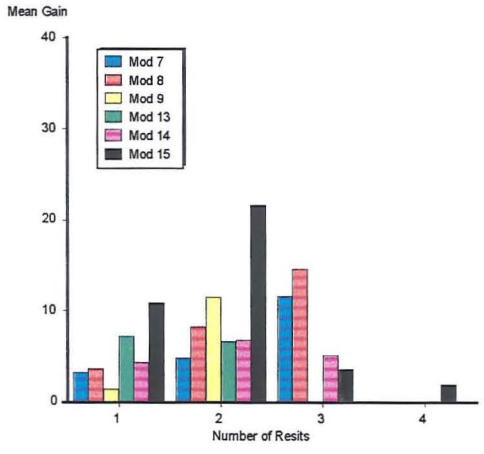


N.B. Module 6 was not re-taken.

From the graph above it is clear that, with the exception of module 4, the greater the number of resits the greater the mark improvement. One might infer that there was some merit to be derived from sitting modules early in the course to allow for more resits. However one could also advance the theory that by waiting until further on in the course, candidates taking the module for the first time would be more mature and would have potentially less to gain from further resits. The gradient of improvement with the number of resits also provides evidence that resitting results in a genuinely higher level of

attainment, and not merely one that derives from 'chance' because of the unreliability of the examining process (both marking and grading).

Figure 5.6: Gains on Resits for Modules 7 to 9 and 13 to 15



There are two observations worthy of note; firstly the gains on these modules are not as great as on the pure mathematics modules and secondly that there is less consistency in gain as the number of resits increases. In particular, gains on module 13 appear to decrease with the number of resits. Whilst a possible explanation might be an inconsistency of grading standards over time, it is difficult to believe this to be the complete story. Module 13 is the most popular for resitting possibly because it is the first optional module chosen by weaker candidates. It might therefore follow that these candidates benefit less from repeated resits than others, but it is an argument that is difficult to sustain in the light of other module results. That the same phenomenon is apparent in module 14 for second and third resits suggests that the explanation may lie in the nature of statistics as a discipline. It might also be compounded by a gender effect (see below).

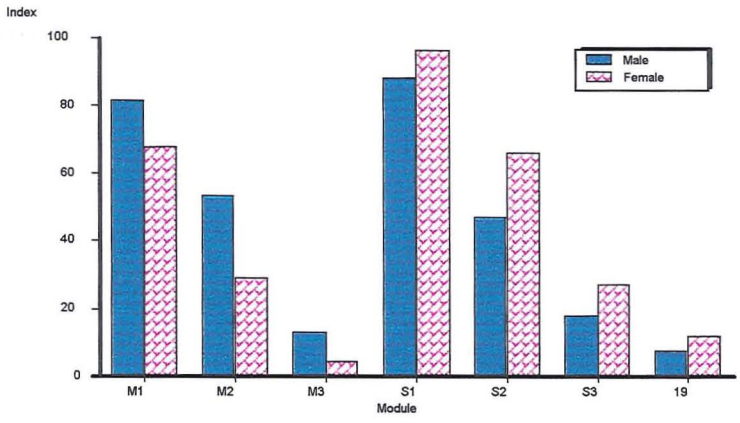
Also of interest was the percentage of the total number taking each module over the various sessions who were certificated in 1994. For example in the summer of 1992, a total of 3617 candidates took the module examination for Pure Mathematics 1. Of these 486 (including Further and X category candidates), 13.4%, requested certification for mathematics in 1994. It might therefore be expected that of the 8645 PM1 candidates in

the summer '94 session, about 1300 candidates would be cashing in their modules for a syllabus award in the summer of 1996. In the same module in the summer of 1993, 48.4% of the candidates requested certification in the summer of 1994. It might therefore be expected that about 4200 of the summer '94 module 1 candidates would end their course the following summer. Given these figures, and the usual 2 year course, the expected proportion of PM1 summer '94 candidates who would request a syllabus grade would be about 35%. In fact it was 7.5%, 648 candidates. Even allowing for an increasing number sitting PM1 in year 11, there would seem to have been a change in entry pattern for module 1 over the four years since the scheme was instituted. Part of the problem stems from the (relatively) large number of centres whose certification candidates in 1992 took all their modules at one sitting. In 1994, there was just one centre! The same analysis could be performed on the other popular modules with similar consequences. It was also possible to calculate how many resit candidates there were likely to be within each module population, but it would seem that the populations are probably too volatile to come to any firm prediction as to the likely composition of any one.

Gender Effect

Of most interest here was the relationship between gender and option choice. Although there was little difference between their average overall grades (5.34 for males against 5.36 for females), the composition of the total entry was two thirds male to one third female, i.e. a ratio of 2:1. However this pattern was not evenly repeated across all modules. This is shown in Figure 5.7 where each module entry is plotted as a proportion of the total entry by gender. Thus the total male entry, which was 2434, was used to index all the male entries for the optional modules e.g. 1981 males entered module M1 so the index was 81.5. The base for the female index was 1189.

Figure 5.7: Indexed Entry by Gender to Optional Modules



What is clear from the above was that relative to their entry, more females, in fact over 96% of females chose S1 and proportionately entered more of their cohort for the statistics options than males. The corollary would be that more males proportionately choose mechanics options and again this is graphically illustrated above. Certainly it was clear that the "third" option for females was more likely to be a statistics module, and for males mechanics though the split between the two disciplines was much more even for this gender. What was clear from the examination of average grades was that those girls who chose mechanics were obtaining higher grades than boys choosing the same option.

Another interesting feature of the gender split was the different resit pattern. In the first two compulsory modules, about 6% more of the male cohort chose to resit at least once (there were very few resits for the later modules so the pattern for both sexes was essentially the same i.e. that of no resits). For M1 the gender resit pattern was very similar, but again for M2 just over 6% (as a proportion of the gender cohort for that module) more males resit. However for the statistics and D & D modules there was very little difference in the resit pattern. A possible explanation is that there were weaker candidates in the male cohort who required to resit, another that they started to take modules earlier giving them not only more resit opportunities but also, especially where maturity might be seen to be a factor, possibly increasing the need for a resit.

Relationship between Modules

One interesting aspect of modular schemes is how each module result correlates with another. The six most popular combinations of modules, i.e. those with more than 100 entries, are shown below together with the grade distribution for each combination. Whilst most candidates take only six modules, it should be remembered that some candidates take more than this minimum number (the average for single subject candidates is 6.1) and thus if they appear in the data for module 7, 8 and 9 say, then they could also appear in the set for candidates taking 7, 8 and 13 say (because their optional modules are 7, 8, 9 and 13), with the higher of their two scores being used for grading.

Table 5.6: Grade Distribution by Module Combination

MODULES	A	B	C	D	E	N	TOTAL
07 08 09	22.5	30.3	21.2	16.1	8.6	1.3	373
07 08 13	27.4	28.8	22.8	13.3	6.4	1.1	1319
07 13 14	23.1	28.3	24.4	14.5	8.4	1.4	1092
13 14 19	14.0	18.4	27.2	26.3	12.3	1.8	114
07 13 19	11.7	26.5	25.1	21.1	14.8	0.9	223
13 14 15	21.6	27.6	22.3	16.8	11.0	0.8	764
Total	23.3	28.1	23.2	15.4	8.8	1.2	3885

These totals are for single subject candidates only, since combinations of subjects are really meaningless for double subject candidates, nearly all of whom will have covered each combination. Not surprisingly the most popular combinations comprise the first statistics and mechanics modules plus the second in either of these two disciplines. It is, however, worth noting that the three statistics modules prove to be far more popular than the three mechanics modules with over twice the number of graded candidates.

The first noteworthy observation is that there are significant differences in the grade distributions for the different sets of modules. Clearly those combinations containing module 19, which is assessed entirely on coursework, are, in general, producing lower numbers of high grades than other coalitions of modules. The reasons for this

dichotomy are unclear; it may be because 100% coursework components attract weaker candidates per se or possibly there is an inherent easiness in this module (which could equate to a high grade threshold).

If we focus attention first of all on the grade A data for the combinations 7, 13 and 14 and 7, 13 and 19, it may be possible to tease out whether there is any evidence for different grading standards. The table below gives the module data for this analysis. Since all marks are in standardised UMS form, direct comparisons can be made.

Table 5.7: Grade A Mean Marks for Two Module Combinations

MODULE	MEAN MARKS	
	Combination 07, 13, 14	**Combination 07, 13, 19**
01	66.0	66.1
02	62.6	61.5
03	60.3	60.7
07	62.9	63.1
13	63.1	62.4
14/19	60.0	59.5

In essence this analysis is using the common modules 1, 2, 3 and 13 as references to the standard on modules 14 and 19. The need for relevance and lack of bias are thus required (Newbould and Massey, 1979). Relevance, given usually by correlation coefficients (see table 5.6) does vary, but given the similarity of results from all the references, may be considered reasonable. Freedom from bias cannot, however, be guaranteed, especially given the nature of the assessment of module 19. Therefore, with these reservations, there are no significant differences (at the 10% level) between the module scores for each of the combinations, including module 14/19. The implication is that grading standards are probably similar across these two latter modules and certainly there is no evidence that coursework offers an easy option. The lack of high grades is thus more likely an effect of a weaker candidature. Whilst there is obviously a need for more in depth analysis to consider the variability between option choices (the subject of chapter 6), whether there is an easier route to an A level grade,

whether there is comparability within this particular modular scheme, preliminary findings suggest that there is perhaps a good deal more comparability than might be expected.

Before leaving this subject until later, a similar analysis for the two most popular combinations, one of which is 7, 13 14, results in the following table:

Table 5.8: Grade A Mean Marks for Two More Module Combinations

MODULE	MEAN MARKS	
	Combination 07, 13, 14	**Combination 07, 08, 13**
01	66.0	66.0
02	62.6	63.0
03	60.3	59.8
07	62.9	64.2
13	63.1	62.4
08/14	60.0	60.6

The results of significance testing here shows that at this grade, for these combinations of modules, there are differences between the results for both module 7 and module 13 (5% significance level, two-tailed test). However they are not easy to interpret since those doing two mechanics modules do better on the first of these, and those doing two statistics modules do better on the first of the statistics options. Since this result is in line with intuition, and since the first three modules show up no significant differences it is probably fair to conclude that there is still little evidence of inequity between module options.

However the focus of this section is on the correlation between modules. Whilst in a traditional A level we would not expect to find perfect correlation between components, especially where coursework was involved, generally the written components would exhibit Pearson correlation values of, typically, 0.7.

The table below gives correlation coefficients for the most popular modules. Here all candidates results have been included, although it is understood that double award

candidates have a more even profile over the modules considered here which will tend to reduce the variance between modules.

Table 5.9: Correlation Coefficients between Modules

1	2	3	4	5	6	7	8	9	13	14	15	19
1	.77	.72	.63	.42	.20	.72	.68	.67	.67	.68	.68	.50
2	1.0	.84	.74	.66	.57	.75	.73	.77	.68	.74	.75	.50
3		1.0	.76	.63	.56	.75	.74	.80	.69	.76	.77	.63
4			1.0	.73	.61	.58	.58	.67	.63	.69	.53	.41
5				1.0	.69	.50	.53	.65	.46	.55	.50	.36
6					1.0	.45	.48	.46	.40	.51	.33	.33
7						1.0	.78	.74	.68	.72	.50	.63
8							1.0	.80	.66	.63	.58	.46
9								1.0	.60	.53	.59	.54
13									1.0	.73	.73	.52
14										1.0	.80	.51
15											1.0	.6
19												1.0

There are a number of obvious points that can be made about these results. Firstly the first three compulsory modules correlate well between themselves, even though module 3 contains an element of coursework. The lowest correlation is found between module 6 and module 1. On consideration this is not particularly surprising because almost all candidates sitting module 6 are double award candidates who would again almost universally have achieved a high A for module 1. Thus the correlation is between a very narrow spread of marks and a very wide mark range. Correlations between these sorts of distributions are mostly meaningless as they reduce to a point/line correlation. (In this context, correlation is a measure of how two sets of observations co-vary. If one of those sets is a point then there can be no 'covariance'). A similar sort of effect is seen with module 5 where there will be slightly more candidates from the single or AS award to introduce a variance on module 1 results.

Again it is interesting to note that module 19 is very uncorrelated with any of the other modules, especially the later papers set for pure mathematics (modules 4, 5, and 6) which contain no coursework. Its assessment regime of four tasks (the best three to count) and a file compiled from work done throughout the course make it a very different type of module which, when compounded with nature of the content of design and discrete mathematics, probably explains the lack of correlation.

Particularly noticeable is the relatively high correlation figures between the first three modules within the disciplines of pure mathematics, mechanics and statistics, all of which are over 0.7. In general, despite the resitting and different sessions at which modules are taken, there is a correlation between marks, which, despite transformation, imply an underlying continuity of standards across like modules. For example the correlation between module 1, which is almost never taken terminally, is subject to resits and probably for each session has the most mixed candidature of any, and module 3, for which none of the aforementioned is true, is 0.72, not particularly high but certainly enough to suggest some internal consistency within and between modules and sessions. This is because the rank order of candidates is determined by their UMS score and if this score is achieved from several sessions (as in module 1), were grading standards across sessions to be very different, then the rank order would be perturbed rather more than a correlation of 0.72 implies.

Module Pairs Analysis

Subject pairs analysis is an analytic device often used to determine whether one subject is, on average, easier than another. Assuming such comparisons are at all valid, there are two further often unwarranted, assumptions made:

i. that the sample population of candidates taking the two subjects is representative of the population of each separate subject,
ii. if two subjects are comparable under a pairs analysis, then for every candidate who gets a better grade on subject A there must be another candidate who exhibits the same differential performance but in favour of subject B. Any divergence leads to the implication of a difference in grading standards.

At A level with unmatched subjects, English and chemistry say, neither of these assumptions can be shown to hold because of the very small numbers who take both examinations. Thus a subject pairs analysis often can (and does) lead to very misleading conclusions.

However, the problems with the above are not so accute within a single scheme of assessment. Therefore, module pairs analysis may be a suitable method of demonstrating comparability within the modules of choice, especially if there is a way of checking that samples are representative. With the analysis below, evaluated on the results of all candidates (because here the greater the number of candidates the more representative the subset of the module pair), while it is to be hoped that assumption (ii) holds, by using a monitor, it is possible to test whether assumption (i) is true by comparing the performance of the subset of candidates from the module pair on a compulsory module with the performance of the population on the same module. In a sense since all candidates take modules 1, 2 and 3 any differences between them will be the same for all candidates. It is with the optional modules that differences can be critical, when a judicious choice could ensure a superior final subject grade were such differences found to be significant.

For the compulsory modules, the mean marks were 57.4 for module 1, 48.9 on module 2 and 43.0 on module 3. These differences are statistically significant ($p < 0.05$), but because they are compulsory for the A level examination, are only of passing interest here. The choice of module 1 scores, rather than module 2 or 3, as a monitor is fairly arbitrary though arguably module 2 might have been better because of its slightly higher correlation with the optional modules. The use of an integral part of the assessment should be sufficient to guarantee greater relevance and less bias than any independent test. It is justified by work reported in chapter 6 where a more rigorous multi-level regression model is applied and where the lack of independence between the first three modules caused considerable problems with the modelling. Hence it is unlikely that any findings would not prove robust were a different compulsory module used instead. In this, somewhat simpler model, module 1 is used to determine whether the population of the sample taking the two given optional modules is representative of the population as a whole[1], the following differences have been found for the popular doublets:

[1] The two-tailed test determines, at the 0.05 level, whether the mean of the sample population i.e. those candidates taking the two modules in question, differs significantly from the population mean taken over all candidates.

Table 5.10: Mean UMS Differences between Modules

Module A	Module B	Mean Difference (A-B)	Number	Module 1 Mean
4	5	4.7	466	66.5
4	7	-6.9	689	66.0
7	8	7.7	2438	60.0
8	9	6.5	980	63.0
7	13	0.2	3347	58.3
13	14	6.8	2723	57.8
13	19	0.2	471	54.9
7	19	-2.4	377	55.7
7	21	7.6	91	66.0
13	21	6.9	83	66.1

In all cases, except the pairings of 7 & 13 and 13 & 19, the difference in the mean values is significant ($p < 0.05$). However, in only one case, the doublet of modules 13 and 14, does the sample not differ significantly from the whole population ($p < 0.05$) as monitored by performance on module 1 and thus assumption 1 does not hold for most of the module pairs. Only in the aforementioned case would it be fair to conclude that module 14 would seem to be somewhat more difficult (by 7 points) than module 13. In all other cases the significant differences indicate that the population choosing the two modules which are being compared, differ in performance from the total population of candidates. Although there appears to be a graduation in difficulty within each of the three disciplines, it would not be sensible to base any firm conclusions on these figures alone. However there is a fairly clear pattern to the results, both the first modules in statistics and mechanics appear of comparable difficulty, and in comparison it may be assumed that the second module in each discipline is comparable. Since the majority of single subject candidates take modules 7, 13 and one other, there would appear, from these results, reasonable comparability in the difficulty for most candidates. Module 4 (almost always taken by further candidates) or module 21, in the usual pattern of module sitting, are probably like the second mechanics or statistics module with which they

equate. Thus although there would appear to be a gradation of difficulty, for most candidates the options together would equate.

Internal Consistency

The usual measure for the internal consistency is Cronbach's coefficient alpha, also known as the reliability coefficient (Cronbach, 1951). It has a maximum value of 1 when the correlations between each pair of variables, in this case modules, is also 1. It can take a negative value when the co-variances between some modules are negative. The defining equation is given below:

$$\alpha = \frac{p}{(p-1)} \frac{\Sigma cov(y_i, y_j)}{var\ y_t}$$

where p = number of modules

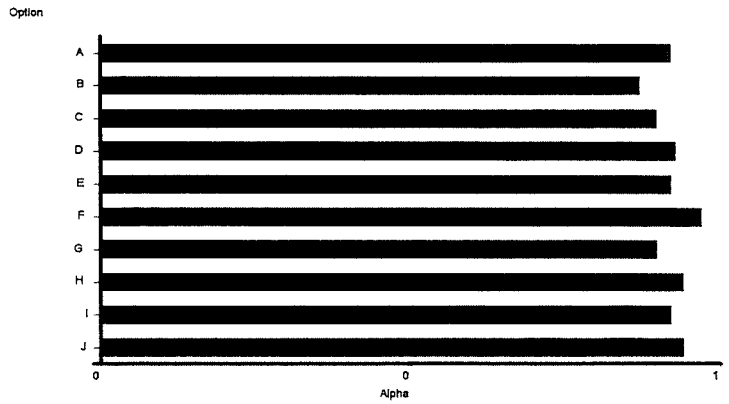
y_i= observed value on ith module

y_t= total observed score (= Σy_i)

Naming the most popular combinations A to J as given in the key below, the variations in the value of coefficient alpha across the options is shown in figure 5.8 whose key is shown below:.

Key:	Module Combination	Key:	Module Combination
A	1, 2, 3, 7, 8, 13	B	1, 2, 3, 4, 7, 8
C	1, 2, 3, 7, 8, 19	D	1, 2, 3, 7, 13 14
E	1, 2, 3, 7, 13, 19	F	1, 2, 3, 4, 13, 14
G	1, 2, 3, 4, 7, 13	H	1, 2, 3, 13, 14, 15
I	1, 2, 3, 13, 14, 19	J	1, 2, 3, 7, 8, 9

Figure 5.8: Coefficient Alpha



The most noticeable feature of the figure is the high value of alpha across all the different combinations of modules, only once dropping below 0.9. This is yet another indicator that there is a high internal consistency in all the module choices, and since modules 1 to 3 are common to all options argues for between option consistency as well. Whilst consistency is strictly a measure of high correlations it does suggest that the rank ordering of candidates is similar across modules which may be indicative of a degree of comparability between them.

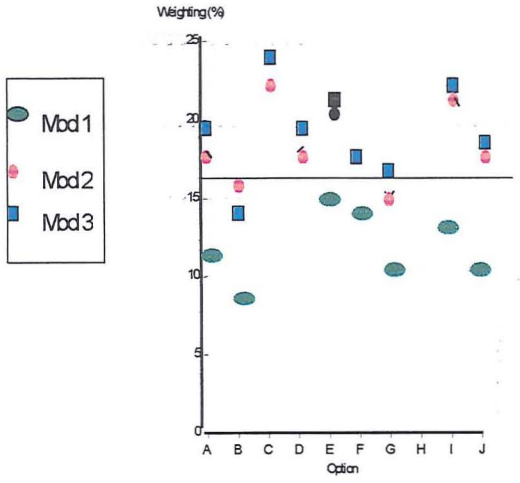
Weighting

The final test, and perhaps one of the most interesting, is that of calculating the achieved weighting of each of the modules in the combinations A to J above. The assessment scheme is based on six equally weighted modules, i.e. each module carries 16.67% of the total marks. However exact weightings are rarely achieved because it is unusual for all parts of an examination to be equally discriminatory. The most useful definition of weighting W (most useful because the sum of the weightings is unity) given below (Fowles, 1974; Adams and Wilmut, 1981) is based on a covariance of part with whole model. With the variable definitions as before, the weighting is given by:

$$W = \frac{cov(y_i , y_t)}{var\ y_t}$$

If a module has a higher than 16.67% weighting it indicates that it has more influence on determining the final outcome than a module weighted at less than the norm. There are a number of reasons why weighting may be lost (or gained). It is commonly seen in tiered examinations where one component, usually coursework, extends over all tiers but for which the other components in each tier have a ceiling grade. Ostensibly the whole mark range is available to all candidates, but in practice, weaker candidates rarely reach the heights and there is a natural ceiling placed on the coursework component which is less than the maximum mark. The same effect would be seen in the context of modular examinations if, for example, one question in a module were wholly inaccessible to candidates and the actual maximum mark was lower than expected. Equally, and perhaps perversely, if a module is very easy, the ability of that module to discriminate between candidates would be lost resulting in a lower than expected weighting. Weighting may therefore be seen as a measure of the proportion that a given module/component contributes to the determination of the final rank order of the candidates. Figure 5.9 illustrates the achieved weightings for the compulsory modules.

Figure 5.9: Achieved Weightings of the First Three Compulsory Modules



Of note is the lower than expected weighting for module 1, for which, while candidates score more highly, the discrimination is low, whilst module 3, in most combinations, contributes more in the determination of the final rank order. It should perhaps be emphasised that weighting of a module is **not** a measure of the proportion of marks contributed by that module to the final score; there is no doubt that as the highest scores are, in general, achieved on module 1, and this contributes more than one sixth of the

total mark. But the weighting of that module, as shown above, is less than the expected sixth. Also it can be seen that the achieved weighting depends on the combination in which the module is found. This is fairly self-evident as clearly the effect of one module on the aggregate is dependent upon the other modules contributing to that aggregate. The weighting of other modules is shown below; each combination is defined as modules 1, 2, 3, a, b, c.

Table 5.11: Weightings Achieved by Optional Modules

Combination			Weighting		
a	b	c	a	b	c
7	8	13	15.1	20.3	15.7
4	7	8	24.0	13.5	23.7
7	8	19	15.1	19.7	7.8
7	13	14	17.0	13.9	19.5
7	13	19	18.9	15.0	10.0
4	13	14	18.0	14.0	18.2
4	7	13	22.0	18.8	19.0
13	14	15	13.3	18.0	19.0
13	14	19	13.2	20.0	9.4
7	8	9	12.2	17.0	23.3

The weighting given to module 8 which varies between 23.7 and 17.0 illustrates the point above clearly. Also worthy of note is the low weighting attributed to module 19. Candidates taking this module have already been shown to be significantly weaker than those taking combinations exclusive of module 19, although the mean score on this module is quite high. Clearly therefore other modules are having more influence on the rank ordering of candidates than the scores achieved on module 19 are. Since this latter is internally assessed coursework, it is perhaps not surprising that there is an apparent mis-match in weightings.

Discussion

This chapter serves to indicate how the modules relate both to each other and how the different variables affect the variation of the module scores. It seems fairly clear from the

analysis above that there is a high degree of internal consistency. This is fundamental to the running of modular schemes, for without this consistency, allowing modules to be taken, and re-taken, over time, would be unacceptable. Since flexibility is to many the reason behind the choice of a modular scheme, the guarantee of continuing and maintained standards is important.

Patterns of behaviour also show that candidates are using the flexibility in time of choosing to take modules throughout the course and the pattern of module sitting indicates that this particular scheme has few candidates who ignore the choices open to them. One result of this is that the facility to resit appears to be used to enhance results, which implies (given the assumption of equivalence of grading standards) that candidates actually show greater evidence of attainment and that therefore they understand the subject content better. While many candidates, in fact three quarters of the cohort, choose to resit a module at least once, results for individual modules reveal that each module is taken only once by the majority of candidates. This finding is linked to the place in the order of assessment of the modules i.e. the earlier in a course that modules are taken, the more likely it is that they will be resat, not only because of maturity, but, probably more importantly, opportunity.

Of those modules for which there was sufficient entry to make any conclusion, it would appear that the 100% coursework module stands out as not conforming to any standard pattern. One conclusion from this must be that modules (ideally each and every one, but certainly those which are interchangeable options) should use the same assessment methodologies. It is sufficiently difficult to achieve balance between modules without this added complication which leads to different cognitive patterns within ostensibly similar tests.

It is now possible to build up a stereotype of the behaviour and performance of the average modular candidate based on the findings of this chapter. Although there is choice available (s)he will probably confine themselves to sitting the minimum number of modules. They will stagger the sessions over which these are taken, probably starting at the end of the first year of a two year course, perhaps taking two or three modules then, with a further one or two in the winter session of the second year.

They will probably resit one or two modules once each, with a consequent gain of about 5 marks per module (slightly more for compulsory modules, and less for optional modules). Should they take more than one resit for any individual module, there will be further improvement over the initial score, but only by one or two marks. Whilst gender would appear to influence module choice, there is little evidence that would support the suggestion that any one set of modules was 'easier' than any other. Whilst the further modules within a discipline prove more difficult, for most candidates who take two modules in one discipline (mechanics or statistics) and one in another, the differential is the same for either choice. This is not ideal, but is a particular problem for mathematics which, of all subjects, is probably the most linear. There is, for example, no pretence that further mathematics is of the same difficulty as mathematics. It is, almost by definition, more difficult. One would not expect to find the same sort of differential in other subjects.

The internal reliability suggests a high degree of consistency within the scheme, but the achieved weighting of each module (i.e. the amount they contribute to the decisions on overall grade) depends largely on the combination in which it finds itself. This is a potential problem for all modular schemes if a strict balance, in terms of means and standard deviations, between modules is not achieved. A factor here may be the conversion from raw to UMS marks, but the care taken with this particular scheme to match raw to UMS scales suggests that this is unlikely, although less well constructed schemes may well find that the prescribed weighting of modules is adversely affected by UMS conversions. Although components in linear schemes may not achieve the desired weighting, the weighting of each component is the same for each candidate. It is therefore essential that this aspect of modular schemes is addressed at the question setting stage.

Whilst it is true there is little positive evidence that any of the previously defined illegitimacies do not exist, there is some negative evidence which goes some way to suggesting that what appear to be inequities are not amenable to generalisation. For example, module pairs analyses cannot be assumed to apply to the whole cohort since, with one exception (modules 13 and 14), the sub-samples taking the pairs are not representative of the population. Resit candidates are improving their results, but there is no evidence that would suggest that this is in any way unfair given the rules of the scheme and the evidence which suggests that the improvement is not just the result of measurement error. In general there is little in this analysis which would counter the

assumption of comparable grading standards although there is clearly a graduation of difficulty in the modules consistent with the nature of the subject and one which should affect all candidates equally.

CHAPTER 6

Keeping One Principal Object in View - Multi-level Modelling

The previous chapter concentrated on the characteristics of the patterns of module-taking exhibited by candidates certificated for single subject mathematics regardless of whether those modules were actually used to calculate the syllabus grade (although for all candidates modules 1, 2 and 3 were compulsory). However it is important not to lose sight of the fundamental problem; namely are all candidates gaining this award exhibiting the same level of attainment? Because we cannot directly address this, we appeal to the underlying assumption of equivalence of grading standards and by looking at the module results attempt to ascertain the truth of this assumption indirectly. Therefore, to approach the topic of 'within subject comparability', the variances between the modules at the different levels at which they are specified are considered. To this end multi-level modelling has been used on what is essentially a multi-response, or multi-variate, problem i.e. six responses per candidate. For this analysis, therefore, the database has been reduced to the six modules per candidate which together formed the total UMS mark from which the syllabus grade was derived. In the case of double subject candidates, an element of choice had been exercised because of the requirement to get the best two grades from the modules available within the syllabus rules. Since in 1994 this was done 'by hand' there was undoubtedly a fairly random element in precisely which mix of modules was used.

Whilst multi-level modelling allows a number of explanatory variables to be investigated, gender for example, it is important to underline the primacy of the inter-module and intra-module relationships to this investigation. In this first section every module is associated with a dummy explanatory variable mod1-mod19. Worthy of note is the statistic that the mean value across all modules and all candidates is 48.93 with an associated variance of 279.56.

Summarised in Table 6.1 below are the modules which contributed to the mathematics results. They are presented in a form which again will allow contrast with the output from the model with the mean UMS mark and variance being given for all modules. It is also worth noting that the number cited as those taking modules 1, 2 or 3 are not identical. Although all three modules are compulsory, because it was decided to retain those

candidates obtaining an 'X' grade, regardless of reason, there are a few who failed to complete the course as well as those who withdrew post-hoc.

Table 6.1: Module Data for Mathematics

MODULE	NUMBER	MEAN/OFFSET	VARIANCE
1	4545	57.4	141.3
2	4523	48.9	284.4
3	4504	43.0	322.4
4	229	48.2	262.9
5	69	51.0	160.8
6	21	54.3	119.9
7	3320	50.8	236.8
8	1920	46.2	306.3
9	453	40.7	436.8
13	3923	50.1	212.3
14	2315	44.4	338.2
15	920	45.6	326.2
19	384	48.8	126.1
21	33	53.0	106.5

The table above shows not only that the highest scores are gained on module 1, but that the greatest spread of results are those for module 9, the third mechanics module for which 453 candidates entered. This module also has the lowest mean mark.

There are a number of ways in which to formulate this problem. The data have already been restricted to six modules per candidate, a decision not unrelated to the requirement of the hardware used to analyse the data. (The dataset is large and memory hungry. Modelling a full set of modules for each candidate requires more than the memory capacity of the machine used). The problem can be considered as one of missing data. Although all candidates have values for the first three modules, subsequently where choice plays a part, essentially the modules which candidates have not taken are missing i.e. there is an incomplete module record for each and every

candidate. Excluded are all those modules which have not counted towards certification in mathematics (even though a candidate may actually have a module score for such modules which may have been included in the calculation for another subject, further mathematics, say). The decision to choose only to include the six modules per candidate adds both verisimilitude and symmetry to the problem and satisfies the demand that only modules which relate to the subject under investigation should be included.

The dataset consists of a number of measurements on each candidate, one for each module taken. Its structure is as illustrated below:

ums	ncn	cand	furth	sex	resit	mod	mod1	mod2	.....mod22	w92	s92...w94
55	999	111	0	0	1	1	1	0	0	0	0 1
48	999	111	0	0	0	2	0	1	0	0	0 0
.....											
.....											
.....											
37	999	111	0	0	0	14	0	0.....1.....0	0	0..... .0	
60	999	112	1	1	0	1	1	0 0	0	1.....0	

etc.

The first six records are for candidate number 111, centre number 999. He is a single subject mathematician who sat module 1 more than once, but the best score of 55 was in the winter of 1994. He sat module 2 once, terminally, gaining 48 marks. His final module was module 14, which he had not resat and on which he gained 37 marks. The next candidate is a female who is also sitting further maths. Her first module, taken once, in the summer of 1992 scored 60 marks.

The variables shown are;

ums	uniform mark score for identified module
ncn	centre number, an identifier
cand	candidate identifier
furth	if 1 then a further maths candidate, else single subject
sex	if 1 then female, else male
resit	if 1 then module resat
modno	module number identifier
mod1..... .....mod22	set to 1 for the module taken, else 0
w92...w94	set to 1 for session in which marks scored, winter 92, summer 92 etc. If all 0 then module sat terminally in the summer of 1994.

Most of the values in the dataset will be zero, so the problem to be investigated is one not only of repeated measures, but also of missing values. Goldstein (1995) has shown

that the best method of investigating such data is to use a multi-variate, multi-level model, and much of the explanation which follows is derived from his book, although the technical justification of the multi-level approach is not repeated here.

The Variance Components Univariate Model

The initial investigation involves an analysis of each module separately and subsets of the full dataset described above were created, one for each module. This essentially removed the missing data problem, because the model only analysed the responses for those candidates who had a response. As will be shown later this is also the only practical approach. The model is known as a variance component model and in this case is specified on two levels - those of candidates within centres. The notation used is for ease of understanding and consistency with chapter 8 which, if a more usual form were used, would necessitate an unwieldy number of subscripts.

Therefore, allowing two levels of variance, (level 1 at candidate level, i, and level 2 at centre level, j), the equation specifying the model contains a constant term and is posited along reasonably familiar lines so that for candidate i in centre j the ums score, y_{ij} is

$$y_{ij} = \alpha_0 + e_{ij}$$

Additionally, the ums score may vary between centres. We can model this by adding an extra term u_j to the equation above. We now have:

$$y_{ij} = \alpha_0 + e_{ij} + u_j$$

with the variance of e given by σ^2_e and of u, σ^2_u. The variance of e is known as the level 1 variance because it is the total variance of scores around the mean for each centre, and the variance of u is the variation between centres. With no other explanatory variables this equation may be written:

$$y_{ij} = \alpha_0 + \beta_{0ij}$$

where α_0 is the fixed part of the model and $\beta_{0ij} = e_{ij} + u_j$ is the random part.

If we now add a dummy variable x_{1ij}, say, which may have a random coefficient γ_{1ij} at level 1 and β_{1j} at level 2. The equation which describes the behaviour of y_{ij} is now given by:

$$y_{ij} = \alpha_0 + \beta_{0ij} + (\alpha_1 + \beta_{1j} + \gamma_{1ij})x_{1ij}$$

with the level 1 variance given by $\sigma^2_e + 2\sigma_{e\gamma} + \sigma^2_\gamma$, where

$$\text{var}(e_{ij}) = \sigma^2_e ; \ \text{var}(\gamma_{1ij}) = \sigma^2_\gamma ; \ \text{covar}(e_{ij}\,\gamma_{1ij}) = 2\sigma_{e\gamma}.$$

and the level 2 variation:

$$\text{var}(u_j) = \sigma^2_u ; \ \text{var}(\beta_{1j}) = \sigma^2_\beta ; \ \text{covar}(u_j\,\beta_{1j}) = 2\sigma_{u\beta}$$

Clearly other dummy variables, x_{nij}, can be added with which are associated fixed, α_n, and random level 1, γ_{nij}, and level 2, β_{nj}, coefficients.

If one defines x_{0ij} as 1 everywhere and the x_{nij}, $n \neq 0$, as 0/1 dummy variables, then it is possible to conflate the above equation such that:

$y_{ij} = \Sigma\lambda_m x_{mij}$, where $0 \leq m \leq n$, and n is the number of dummy variables,

and

$$\lambda_m = \alpha_m + \beta_{mj} + \gamma_{mij}, \quad m > 0$$
$$\lambda_0 = \alpha_0 + \beta_{0ij} ;$$
$$\beta_{0ij} = e_{ij} + u_j$$

This notation is more in line with the convention used by Goldstein (1995).

Using this model, each of the modules was investigated in turn.

Module 1

All explanatory variables were initially included in the modelling for all the modules, and removed if they proved to be statistically insignificant. For module 1, with 4545 level 1 units and 211 level 2 units, the explanatory variables which proved significant were:

Fixed	**Estimate (se)**	
intercept	58.31	(0.76)
resit	-9.34	(0.57)
furth	5.35	(0.43)
w92	-6.13	(1.01)
s92	-2.65	(0.97)
w93	1.81	(0.75)

These estimates show that whilst there was no significant difference in results from sessions in the previous summer and winter, those who sat module 1 early in the course were, in general, obtaining lower marks than those sitting later on. Also it can be seen that those sitting additional modules, i.e. further mathematics students, were obtaining half a grade higher than those who were only entering single mathematics.

Perhaps of most interest is the contrast for the resit dummy. This suggests that candidates for whom this module score was the result of a resit were getting nearly a grade lower than their counterparts who sat the module only once. Since we know from previous investigations that in almost every case where a module is resat, the score is enhanced, it is clear that candidates although benefitting from the resit regime are weaker than those who sat the module only once. Allied to the session results, it is possible to conclude that if candidates sat the first module very early in the course (probably year 11), and this was their best result even if they re-sat, they would be obtaining lower than average results but were probably more capable than those whose best result came from a re-sit. Gender has rarely been a significant factor in any of these module results.

The random part of the model produced the following results:

Random **Level 2 (School)**

	cons	resit	w92
cons	23.30 (3.10)		
resit	2.33 (2.76)	19.50 (4.61)	
w92	21.86 (3.18)	-6.58 (3.12)	9.57 (3.79)

Level 1 (Candidate)

	cons
cons	59.75 (1.63)
resit	32.00 (2.40)

The variance at level 1 (given by 59.75 + 2*32.0, i.e. 123.75 if resits are included, though only 59.75 if not) is approximately twice that at level 2, with only the co-variance term between resit and 'cons' not proving to be of significance. The level 2 results show that there is significant variation in responses for both the winter '92 session and resits. This suggests that perhaps the decision on resits and whether year 11 children should take modules are very much dependent on school policy with little random variation detected at the candidate level. Likelihood ratios for all these models are given in appendix F.

Module 2

The second of the compulsory modules produced the following data:

Fixed	Estimate (se)
cons	47.62 (0.86)
resit	-10.42 (0.77)
furth	8.87 (0.70)
w93	6.30 (1.60)
s93	6.92 (0.89)
w94	3.44 (0.65)

The fixed part of the model shows that the later sessions have become significant and that candidates who are sitting this module early in the course, but not in year 11, are gaining higher marks than those who sit the module terminally. The resit factor is again significant and the performance gap between those sitting further mathematics and those who do not is widening with the more difficult module.

The significant random elements in the model are

Random	Level 2			
	cons	resit	s93	w94
cons	62.84 (12.28)			
resit	-28.02 (9.32)	41.15 (10.08)		
s93	-33.51 (10.46)	35.00 (9.29)	34.89 (11.5)	
w94	-15.27 (7.31)	19.33 (6.26)	12.03 (6.79)	6.49 (5.77)

Level 1

	cons
cons	137.60 (3.95)
resit	34.51 (4.07)

Although the level 1 resit - intercept covariance (which increases the level 1 variance from 137.6 to 206.6) is still important, there are again more variables adding to the level 2 variation than found at level 1. This is specifically true of the session indicators. Whilst their level 1 variation is not significant, the fact that session variance is found at centre level probably indicates that choice of when to sit modules is more likely to be a centre decision than the individual candidate.

Module 3

The fixed part of the model estimates the significant offsets as shown

Fixed	Estimate (se)
cons	37.35 (0.70)
furth	13.37 (0.90)
s93	11.21 (1.04)
w94	11.18 (0.94)

A far lower overall estimate and the widening of the gap between further and non-further candidates betokens a more difficult module and the session offset was probably the result of predominantly those better mathematicians possibly taking the further AS option (not included in 'furth'). The lack of significance of the resit variable shows that, as expected, it is a rarely retaken module.

The random parameters are as follows:

Random	Level 2			
	cons	resit	s93	w94
cons	71.98 (9.62)			
resit	-35.81(10.95)	50.23 (16.05)		
s93	-34.83 (8.85)	33.02 (10.33)	19.7 (9.65)	
w94	-39.92 (9.96)	38.01 (10.76)	27.15 (9.26)	39.13 (13.05)

	Level 1
	cons
cons	205.9 (4.51)

At level 1, only the residual variance about the mean is significant and it explains almost two thirds of the total variation. This is because choices are restricted for the later modules and the school influence is less pronounced. Still there is much level 2 variation, not just in the constant term, but also in the between school variation in the penultimate session term and in the resit factor. Additionally the negative co-variances between the random coefficients of the factors suggest that the between school variations in the results obtained from those candidates who have taken the module in the winter of '94 are very different from the overall differences in the module 3 results between schools.

Modules 4, 5, and 6

The results of these three optional modules are presented together for comparison purposes. The fixed part of the model estimates the following as significant::

Fixed	**Estimate (se)**
Module 4	
cons	44.05 (1.8)
furth	7.6 (1.67)
w94	4.92 (1.95)
Module 5	
cons	43.46 (3.88)
sex	8.76 (4.10)
Module 6	
cons	26.00 (8.30)
furth	28.47 (8.61)

The variables with significant estimates for modules 4 to 6 are given above. For module 5 there is no significant result for the further dummy. This would suggest that of the very few sitting this module who did not take a further option, they were performing to the same level as those who did or that the data were too few to detect any effect. Whilst module 6 is very much out of line with the other modules this is certainly due the fact that not only was there a very small entry but that there was only one non-further candidate who performed poorly. Because it is the last pure module other factors would be zero.

The random elements are given by the intercept variances below:

Random	
Module 4	
Level 2	68.47
Level 1	89.5

In each case the standard error is < 0.005.

Module 5	
Level 2	123.2 (38.74)
Level 1	51.4 (12.29)

Module 6

Level 2	16.76 (20.83)
Level 1	52.07 (20.31)

In all three cases the significant random variation can be attributed to that which exists between candidates or between schools as no other variables add to the level 1 or level 2 variances. What is of interest is the amount of the total variance which exists at either level. For module 4 the level 1 variance is somewhat larger than that at level 2: for module 5 the level 2 variance is over twice that at level 1: and for module 6 the level 2 variance is not significant. Since none of these modules contain coursework, the obvious reason for variation in standards between schools, it might be inferred that the teaching input to module 5 is more significant than for other modules.

In an attempt to gain more understanding of the variation at the two levels all the module analyses (except those for modules 1 to 3 for which all results had already been analysed for the '94 cohort because of their compulsory inclusion in the assessment) were repeated but with no restrictions on the number taken by each candidate. For the later modules, nearly all of whose uptake would count towards the further mathematics grade, an enhanced the number of results could be included.

Mechanics Modules 7-9

The mechanics modules taken by single subject candidates give the following estimates for the fixed part of the model:

Fixed	**Estimate (se)**
Module 7	
cons	47.11 (0.79)
furth	6.25 (0.85)
sex	-1.78 (0.54)
w92	-11.82 (5.04)
s92	7.68 (2.02)
w93	9.42 (1.25)
s93	7.30 (0.71)
w94	5.56 (0.73)
resit	-7.89 (0.73)

Module 8

cons	42.59 (0.89)
furth	10.61 (1.34)
s93	5.03 (1.49)
w94	5.66 (1.02)
resit	-4.24 (0.97)

Module 9

cons	36.81 (2.00)
furth	14.62 (2.78)

Again for the earlier modules more factors are significant. For module 7 all sessions produce significant estimates, and indeed one might conclude that the winter 93 session module 7 examinations were somewhat easier than others. Certainly if all the module examinations were set to the same standard then the session which produced the best result could be said to be indicating an optimum sequence. For module 8 the resit offset was diminished. Once advance to module 9 had been made, session was not a significant factor, only those taking further modules producing significant offsets.

The random parts of the models produce the following:

Random

Module 7

Level 2

	cons	resit	w92
cons	44.01(6.48)		
resit	-10.39(5.65)	23.77 (7.9)	
w92	109.3 (35.7)	-84.13 (33.65)	288.7 (174)

Level 1

	cons
cons	125.8 (3.96)
resit	29.74(4.59)

Again for modules 8 and 9 the variances are those for the interceprt only:

Module 8

Level 2	52.22 (9.25)
Level 1	229.2 (7.69)

Module 9

Level 2	140.7 (37.72)
Level 1	256.8 (18.81)

Only for module 7 is there variance explained by anything other than the constant term. Here too there is a greater variance at level 2 than at level 1, although not all due to the 'CONS' variable. Indeed the between centre variance for those sitting the module in year 11 appears to explain much of the observed variance. There is also a significant co-variance term between W92 and 'CONS'.

Statistics Modules 13-15

The fixed part of the model produces the following estimates for the statistics modules:

Fixed	**Estimate (se)**
Module 13	
cons	47.42 (0.89)
sex	1.41 (0.46)
resit	-8.11 (0.59)
s93	3.94 (0.89)
s92	5.72 (1.68)
w93	7.18 (0.99)
furth	6.11 (0.73)
w94	4.60 (0.73)
Module 14	
cons	39.57 (0.93)
sex	1.78 (0.76)
resit	-8.30 (0.91)
s93	6.11 (1.20)
furth	10.97 (1.18)
w94	6.70 (0.84)
Module 15	
cons	42.07 (1.16)
sex	3.37 (1.22)
furth	12.76 (2.07)
w94	6.82 (2.00)

The rather familiar pattern emerges from these three statistics modules. Perhaps somewhat more of the factors appear significant on module 15 than might have been

expected, but nevertheless there is considerable agreement, especially on the sessions which produce the best results. Indeed there is a clear sequence of offsets; a maximum value is found for the winter '93 session for module 13, module 14 has the best results when candidates sat in either summer '93 or winter 94, and finally for module 15 the best results were achieved in the winter of '94. The resit factor was similar in the two early modules, not featuring latterly. There is a small gender effect which shows that girls have slightly better results than boys on the statistics modules.

The random elements produce the following output:

Random

Module 13

Level 2

	cons	resit	s93	w94
cons	65.48 (10.9)			
resit	-12.88 (5.96)	12.4 (5.13)		
s93	-27.63 (8.99)	17.32 (5.75)	36.3 (10.66)	
w94	-19.91 (7.21)	8.64 (4.16)	18.97 (7.21)	17.82 (7.16)

Level 1

	cons
cons	109.50 (3.70)
resit	26.28 (3.32)

Module 14

Level 2	66.29 (10.72)
Level 1	211.1 (7.73)

Module 15

Level 2	45.06 (0.00)
Level 1	242.7 (0.00)

For all three of these modules the level 1 residual variance explains most of the variation although for module 13 the session is significant. The one oddity is the lack of a standard error associated with the variance estimates for module 15, probably because there are too few candidates to make an estimate. Otherwise the pattern of the random variation is much as before.

Modules 19 and 21

Although these two modules are discussed together they are very different on content. However their method of assessment, i.e. internally assessed coursework, is the same. This is no longer allowed under SCAA rules and thus their untypicality may not be found for later modules assessing the same components.

The fixed part of the model is given by:

Fixed	**Estimate (se)**
Module 19	
cons	45.64 (1.22)
sex	3.55 (1.10)
furth	7.80 (1.84)
Module 21	
cons	49.94 (4.01)

Predictably few of the factors appear significant, and gender has little effect. One might perhaps expect this to be rather stronger given the received wisdom of girls superiority in coursework. Only in module 19 is 'further' a factor of account, and almost certainly says more about the composition of the candidature for these modules than something intrinsic to the modules which might be of fundamental importance.

The random elements are:

Random	
Module 19	
Level 2	39.26 (10.47)
Level 1	74.17 (5.7)
Module 21	
Level 2	No Significant Result
Level 1	78.48 (23.74)

It might be expected that somewhat more variance would be at school level given the nature of the assessment, but it only explains just over a third of the variance for module 19 and considerably less, somewhat under 15% for module 21. Neither session nor resit variables are of help in explaining the random variation in the scores.

Multivariate Random Coefficients Model

An obvious extension of the above approach to modelling the modules individually, and the effect of several interactive variables is to consider modelling the whole dataset together. This would produce a three level model, with level 1 being the module level, level 2 candidate and level 3 school.

One of the obvious problems with this approach is that there is a considerable amount of missing data. Each candidate only produces responses for six of the 22 dummy variables, which means that over two thirds of the data is effectively missing. Also, only six out of 22 of the measures are repeated, with only three of the measures (modules 1 - 3) common to all candidates. In the dataset there are 27164 level 1 units, the total number of module responses, 4545 level 2 units (the number of candidates, approximately one sixth of the number of records) and 210 level 3 units (the number of different centres).

We can define the multi-variate model by extending the analysis above to include 3 levels of analysis. The basic model, for module k and candidate i in centre j, is now

$$y_{kij} = (\alpha_{k0} + \beta_{k0ij})\, x_{k0ij}$$

where α_{k0} is the mean value of the response for module k, and β_{k0ij} is the random variation of this response such that $\beta_{k0ij} = e_{kij} + u_{kj}$. The candidate level 2 variance of e is given by σ^2_{ek} and the centre level 3 variance of u, σ^2_{uk} for each module k. We can again conflate this such that

$$y_{kij} = \Sigma\lambda_{p0}\, x_{p0ij}, \quad \text{where } 1 \le p \le q, \text{ and q is the number of modules,}$$

and

$$x_{p0ij} = 1 \text{ when } p = k, \text{ 0 otherwise}$$
$$\lambda_{k0} = \alpha_{k0} + \beta_{k0ij} \text{ ;}$$
$$\beta_{k0ij} = e_{kij} + u_{kj}$$

If we now add the dummy variables together with their random components as before then we have:

$$y_{kij} = \Sigma\lambda_{p0}\, x_{p0ij} + \Sigma\lambda_{km}\, x_{kmij}, \quad \text{where } 1 \leq m \leq n, \text{ and n is the number of dummy variables,}$$

and

$$\lambda_{km} = \alpha_{km} + \beta_{kmj} + \gamma_{kmij}, \quad m > 0$$

Random variables are only defined at level 2, candidate level, and centre level 3.

N.B. For convenience Pure Mathematics modules 1 to 6 are designated mod1-mod6, Mechanics modules 1-6 are named as mod7-mod12, Statistics modules mod13-mod18, Decision and Discrete Mathematics mod19 and finally Numerical Analysis mod21. The variables W94, S94 etc denote the session at which the module was taken, and RESIT whether the score was the result of a resit. FURTH is the variable which shows that the candidate also took further mathematics. SEX is the gender variable, which rarely has a value because it turned out to be insignificant in virtually all the analyses.

Attempts to include all the variables in the model met with little success. There are probably several reasons for this, but there are two major, not unrelated reasons, why this should be so. The first is the sheer numbers of missing values. Even restricting the module choice to that which is the most popular combination (modules 7, 8 and 13) for which there are over 1300 candidates still implies missing values for at least one of these modules for over 3000 candidates.

The other obvious source of problems lies in the high correlation between some modules. Although they are separate observations, in modelling terms they appear to be linearly related and are therefore seen as not independent. An unconditional regression analysis certainly suggests that only three of the module scores are needed to explain as much of the variance as is possible with the current data.

Using just three module scores, the software produced a convergent solution such as is shown below:

The fixed variables are estimated as:

Fixed	Estimate (se)
mod1	53.19 (0.58)
mod3	39.03 (0.68)
mod9	32.82 (1.21)
resit	-3.47 (0.28)
w94	2.80 (0.36)
s93	3.90 (0.38)
furth	7.24 (0.50)
w93	3.13 (0.50)
s92	-1.77 (0.77)

and the random

Random	Level 3		
	mod1	mod3	mod9
mod1	32.22 (3.97)		
mod3	44.75 (5.67)	73.83 (9.20)	
mod9	69.97 (9.40)	102.40 (14.21)	165.4 (26.68)
	Level 2		
	mod1	mod3	mod9
mod1	93.41 (2.00)		
mod3	88.72 (2.59)	226.9 (4.89)	
mod9	92.97 (5.88)	199.5 (9.40)	

Whilst this may not give much insight into inter-module comparison since only three are included, it does have one or two interesting features, not least that the modules producing a convergent solution exhibit the widest spread of distributional characteristics of those available. The co-variances between the modules are high at level 3, but lower at candidate level implying that there is rather more similarity between candidates at the same institution than there is over all. The results also indicate that resit candidates are 3 to 4 UMS points below those who do not resit. Also those candidates who take a very early module are probably doing themselves a disservice.

Since it initially appeared desirable to attempt to include all variables in the analysis, other methods of overcoming the linear dependencies detected by the model were tried. The results for modules 1, 2 and 3 were averaged and used as a single dummy variable, but this seemed to enhance the linearity, probably by performing a smoothing function which actually increased correlations between this average and other modules.

Secondly, all the data were used (not just the six modules which counted towards the syllabus grade). This caused the model to crash even earlier in its iterations, probably for similar reasons to those advanced in the paragraph above.

The failure of the model to detect sufficient variation in the scores of individual candidates at module level may be due to model inadequacy or too few data, but one could also conclude that it is the data which insufficiently discriminate to allow the model to perform efficiently. This being so, then the corollary is that little extra is learnt about a candidate's performance once a minimum number of modules have been taken. For mathematics the optimum number indicated by the analysis would appear to be between three and four.

Whilst there are sound educational reasons for candidates having knowledge of the parts of the syllabus examined in each of the six chosen modules, it may be that achievement in the subject can be demonstrated by testing just part of that knowledge providing correlations are reasonably high. Whether this conclusion can be extended to other subjects is problematical, but given similar conditions of rank ordering between achievement in each module, there is no obvious reason why it should not follow. That being so, there may be insufficient reason for insisting on a six module regime when achievement can be amply demonstrated by the taking of four. Conversely, this method of analysis would be eminently suitable for those subjects which employ a number of assessment methods and where correlations between modules are low.

Discussion

A fairly consistent pattern has emerged from the modules, most could perhaps have been guessed at intuitively, but the data here are, for the most part, unequivocal. The messages they send are that on the first modules in each discipline, early sessions lead to the best results (probably because only the very weakest candidates would choose to resit beyond a certain point); that on these modules too taking the resit variable into account helps to explain more of the variance, and entirely consistently further mathematics candidates gain higher marks on these same modules, the superiority falling off for later assessments. Gender is almost never significant either as a fixed parameter or as part of the random variation.

The analysis was repeated for modules 4 to 21 with all the results included i.e. candidates were no longer restricted to 6 modules (all candidates had been included in the analysis for the compulsory modules). Obviously the majority of additions were to further candidates and this variable assumed an importance not generally seen in the restricted analysis. Often the resit variable would also become significant in determining the source of variation.

In general with an increase in numbers came an increase in average score (because the better further candidates were now being included) and in the variance of that score.

Modules 4 and 5 had slightly different estimates for the intercept, but both also now resulted in an estimate for the resit variable of -4.6 and -7.6 respectively, with the further dummy also showing significance for module 5 estimated at 13.58. Module 6 now produced a base estimate of 49.55, a far more realistic value. All estimates of the level 2 and 1 random variations increased approximately doubling in size. The negative co-variance between further and 'CONS', found in many other results most notably at level 2, explained much of the variance of both module 4 and 5. For module 6, the only random variation of significance at both level 1 and 2 was again 'CONS'.

Modules 7, 8 and 9 all had higher mean marks, although the variance for module 9 was reduced. Little difference was observed in the parameter estimates for module 7 and 8, but additional variables of resit and 'W94' produced significant, though small estimates. The estimates for the random variation again showed that for all three modules the negative co-variance between further and 'CONS' was appreciable at both levels of estimate. For module 8 the resit variable also proved to have significant variation at level 1. It is interesting that there is this negative correlation between the base (i.e. non-further candidates) and those taking the further option. This implies that the better the further candidates score, the worse the single subject candidates will score, i.e. the discrimination between the two types of candidate greatly increases with the later modules.

Again, for modules 13, 14 and 15, there are slightly higher mean scores, but little difference is observed in the estimates of the variables from the multi-level modelling. The most notable addition to the random variation at both levels is the further/'CONS interaction for all three modules. For all modules the variance in the further indicator is

significant at level 2, suggesting that the variation in further candidates' results is greater between centres than between candidates.

Little difference is found in the results for modules 19 and 21 except the estimate for the random variation in 'CONS' at level 2 for module 21 becomes significant, equalling the level 1 variation. This suggests, as might be expected with a coursework module, that much of the random variation in the scores is indeed at centre level. It is slightly surprising that the result was not significant with the restricted dataset, although the numbers tripled when all module takers were included.

The multi-level modelling brings a different insight into the working of modular schemes. Initially results suggest that for this subject at least an optimum number of modules might be four rather than the six at presently used. Such a finding could have implications for other subjects whose components/modules are highly correlated. However, the analysis presented does not imply which four, if only four are to be chosen, should be used, and the educational argument that the content and skills covered by each module are, to some extent, independent and therefore need to be assessed separately, is irrefutable. Rearrangement of the content of six modules into four is a possiblility and one that might find favour especially amongst those who consider six module tests to be 'overexamining'. It would also fit in with the two examination sessions a year so that the syllabus could be broken down into four teaching and assessment units.

For all modules considerable level 2 variation was found and for the early modules, the session at which the test was taken proved of significance. The variance of the resit factor at centre level leads to the conclusion that this is often a centre level rather than candidate level decision. Not surprisingly the further estimate (i.e. those candidates also taking further mathematics) added considerably to the base score, but with the resit having a negative estimate there is a clear implication that it is the weaker candidates who choose to resit, and that their scores still lag behind those who do not resit. In fact it is important to stress that an improvement in module score due to a resit does not imply any 'unfairness'. This would only be relevant, given the stated position of the author, if there was an element of luck, probably due to measurement error, which is not the case with mathematics. The very nature of choice inevitably will lead to some candidates sitting module examinations rather early given their ability, and the additional impetus

given by the feedback and experience from the first sitting, plus the re-enforcing of content knowledge endowed by the resit cannot be under-estimated. Even strong candidates may resit to 'get a better A' but the evidence suggests that they do not. Whether candidates are penalised by having to pay for a resit seems (anecdotally) to vary from school to school and, apart from registering that it may be a factor in determining who avails themselves of the opportunity to resit, is mainly conjecture.

That many of the variables do produce significant estimates, and contribute to the understanding of the examination process, has been demonstrated. Precisely how this knowledge can be adapted to enhance the effectiveness of the modular examination experience is not clear, even if it is possible. However few of the findings can be said to cast doubt on either the grading standards across modules or the effectiveness of modular schemes in the assessment of ability, at least in mathematics. Nor is there any evidence of construct irrelevant variance, since all variance which was found is legitimate in terms of the set syllabus.

CHAPTER 7

But Much Yet Remains to be Said - Question Paper Analysis

One of the major differences between conventional and linear schemes has been in the form that the written examinations have taken. Though not universally true, it is in general the case that for A level certification, module written papers are more numerous than their linear counterparts and of shorter duration. This is the case with MEI mathematics and its traditional comparator, OCSEB Mathematics (syllabus number 9650).

Essentially the research question considered in this chapter is whether construct under-representation can be detected when the modular scheme is compared with the linear scheme which it is replacing. If the candidates demonstrate the same knowledge and skills in both linear and modular examinations irrespective of the structure and length of the question papers then one of the potential threats to validity, and thus comparability, may be discounted.

This particular problem has proved especially intractable. Where Bloom's taxonomy and later variants appear in every syllabus under the heading of assessment objectives, there is little attempt to categorise questions in quite the same way (although with the advent of the National Curriculum, question paper grids have appeared in connection with GCSE written papers). Most A level question papers are a matter for the experience and expertise of the paper setters with little systematic attempt to define in educational or cognitive terms what each mark, or set of marks is actually given for. In fact it might be asking the impossible to expect mathematics examiners to say which marks are awarded for, say, skill and which for knowledge when they appear inseparable within the context of individual questions.

Such reliance upon the collective expertise of examiners may work in practice, but is of little help when unravelling the reasons why some questions are hard and some easy, especially when there is often less than universal agreement on which is which! Identification of the source of difficulty, whether it be context, structure or content is especially difficult in comparative terms and at Advanced level where some skills, which may cause difficulty at a lower level, may be assumed e.g. reading and writing. Indeed knowledge of 'subject jargon' is part of the requirement at A level.

What is also of some concern is that most of the literature applies strictly to multiple choice questions, and palpably the type of questions found in A level mathematics papers, whether linear or modular, are more easily characterised as 'free response'. However question choice does arise to complicate the issue, though it is not intended here to investigate specifically such aspects of the linear scheme although it is one of a number of reasons why modular schemes may be inherently more reliable than linear schemes.

Increasingly, modular questions are set with a paper 'design threshold' in mind - at the crudest level there is the intention that the 'a' boundary would be found at the 80% mark, the 'b' boundary at the point where 70% of the raw marks are scored and so on. This is to reduce inconsistencies when conversions to the conventional UMS mark scale is made. In other words mark schemes (written in conjunction with the question papers) are written with expectations of minimum performance at each grade. These expectations are based on the experiences of the question setting team gained over many years of examining.

The MEI philosophy is somewhat different. These papers are set with the intention that 100% of the marks will be accessible to an 'a' grade candidate (keeping to the convention that lower case letters denote module grades), 75% to a 'c' candidate and 50% to an 'e' candidate, and it is expected that candidates will, in general, gain at least 75% of the accessible marks. This gives design thresholds for each module of 75% for grade 'a', 56% for 'c' and 38% for 'e'. Though these ideas may be in the mind of a linear examiner, it is never explicitly stated.

A recent literature review (Hughes, 1995) outlines the work which has been done regarding the estimate of question difficulty, much of which is based on response. One of the best known, the SOLO taxonomy[1] (Biggs and Collis, 1982), is based on levels of response (which for 16+ students would be expected to be 'extended abstract') and is more about the different types of response elicited by the same question rather than question demand per se. However, they acknowledge that the 'nature of the task can affect level of response' (ibid. p 178) without being more

[1] The SOLO taxonomy is a categorisation of question response on five levels from pre-structural (an irrelevant response) to extended abstract (two or more interacting concepts).

explicit. Similarly, even if the outcome space from each scheme were quantifiable, there is no proof of any correlation between such space and question difficulty.

There are also other factors which need to be considered regarding question demand. The unquantifiable effect of, for example, question familiarity, of time pressure and of personal preference (ibid.) which may affect the attainment of candidates in unpredictable ways means that empirical evidence of question demand can and does produce somewhat different, and unexpected, outcomes from that which may derive from analytical considerations. They are part of the explanation as to why examiners sometimes get it wrong!

Recent inter-Board/Group comparability studies have relied on defining demand from a set of factors initially postulated by Pollitt et al (1985). Although devised from work on Scottish 15 year olds it has proved sufficiently robust to be used in wider contexts. The authors postulated three distinct task facets: subject or concept difficulty; process difficulty; question or stimulus difficulty. The work centred on five subject areas, one of which was mathematics, and described a number of factors within each facet which was found to affect question difficulty. Whilst generalised, there was also specification within each subject area. For mathematics, subject difficulty was affected by the familiarity of the concept and the degree to which notation was removed. Regarding process difficulty (often exemplified by the examination mark schemes) the most influential determining factors appeared to be the amount of generalisation required, the difficulty in recognising familiar problems and applying known solutions, the problems found in deriving an underlying principle and in devising strategies to solve an unfamiliar problem. Question difficulty itself centred on the number of steps involved in any solution, the number of arithmetic calculations which could lead to computational errors, the number of different ways in which a correct answer could be calculated, the 'cues' contained in any question which can help candidates to devise their own strategies. The type of question structure which would lead to the correct answer was also found to be an influential factor in the determination of question difficulty, and was found to apply in every subject area considered.

Although many of the factors are pertinent to questions within both schemes i.e. there is no innate feature of either of linear or modular schemes which will unduly influence question difficulty, there are, however, some exceptions. Within the

subject difficulty category lies the concept of 'degree of familiarity' which may be different within the two schemes, even though the subject content is essentially the same. Whilst there is a focusing effect with the modules, the reinforcing of earlier learning objectives later in the curriculum which is usual for many linear schemes, would appear to have a role in emphasising familiarity with certain principles and techniques. In a similar fashion, although most of the subsets concerned with process difficulty could equally affect either scheme, there may be a difference in the cumulative difficulty because of the inclusion of somewhat longer questions in the linear scheme. Such reasoning might imply that there could well be a greater variety and number of mathematical operations included in the obtaining of solutions to linear questions. The devising of strategies also becomes more relevant with more complex questions. However, the unfamiliarity (or otherwise) of the question might apply equally within both schemes of assessment. Also there is little doubt that the greater structuring of questions, a declared feature of the particular modular scheme under consideration, does have some influence on question difficulty.

Recent inter-Board/Group comparability studies have relied on defining demand from these factors, customised for the subject under review, as applied to the different papers presented to them. The procedure relies on the expertise of scrutineers who are called upon to make subjective judgments, based on the pre-defined factors. This remains a reasonably successful approach provided that the structure and question type of all the papers are similar. However, on the only occasion when scrutineers were asked to compare linear question papers with those for modular schemes (SCAA, 1995) the results were inconclusive.

Nevertheless, the suspicion still remains that, in some ill-defined way, the type of questions appearing in modular papers are easier than those from the linear schemes. Certainly modular question papers for MEI are shorter and the questions are deliberately structured - far more than for the comparable linear scheme. However the accessibility of the initial part of the question is no guide to the accessibility of the final part. This final part of the question may be in the form of a rider to the substantive part of the question or merely the final part of the structure but observation and anecdotal evidence shows it is rarely answered correctly and often not even attempted by those who cannot manage the initial parts. In the linear

scheme where the parts of any question may be independent, no such "tenth mark" syndrome exists and accessibility may be evenly spread across the distinct parts.

To compare the demands of two question papers of three hours' duration and six of one hour's duration each is immensely complicated. The wider focus and the scope of longer papers, based as they are on the whole curriculum would suggest that the level of demand is much greater. However, what is of interest has to be the totality of the testing in modular schemes despite their separateness, not the individual elements. Even if the questions are identical in the two schemes, it is impossible to quantify the differences in demand which arise simply because of the scheme of assessment. However it is what is expected of awarders when they choose boundaries and is one of the reasons for the assumption made in this thesis for it puts the equating of demand firmly in the realm of judgement.

For any written examination paper, whilst clearly the total marks available for each question (irrespective of context) are an indication of the examiner's assessment of the question difficulty, and the marks achieved by the candidates an indicator of the actual difficulty, the problem remains of linking any two papers on a common scale so that comparisons can be made.

Methodology

The analysis of question difficulty has been split into three sections roughly approximating to the three facets described by Pollitt et al (1985). Firstly there is a detailed review of the two syllabuses under consideration. This is intended to highlight any differences in subject content although it is impossible even to attempt to quantify the degree of familiarity with the subject. There are a number of conflicting opinions regarding this issue. Firstly it is felt that delivery of a test at the point of teaching, which a modular scheme can accomplish far more effectively than any linear examination regime, focuses effort in a narrow subject area - to the advantage of the candidate. Conversely it may be that the reinforcing of much of the syllabus through constant use and repetition (far more common in linear schemes) increases familiarity and thus facility. It is a debate that is unlikely to produce a winner. This dilemma is clearly a source of illegitimate variability and one that can, at least partially, be resolved by awarders since the level of attainment shown for a given grade should be the same, irrespective of origin, once the the demand of the

question papers has been taken into account. If, at the and of all the analyses, it can be shown that there is a gap between expectations and outcome, one of the reasons may be the testing regime.

The second section outlines the structure of the question papers and their differences, considering the syllabus coverage, generally thought to be better in modular schemes. Mark schemes are consulted for the apportionment of the various marks and a value assigned to each question on the basis of these marks and the time taken to complete the paper. One problem which arises is the variety of choice of modules which is open to some candidates. In order to make the task manageable it was decided to consider only those options which were most popular. Thus certain of the comparisons may seem somewhat slanted, given the relative length of both syllabus and question paper.

Finally some analyses are conducted on the question performance of 100 candidates per component/module chosen at random. For the modular scheme no candidate appears in more than one option, although the linear candidates are common to both components. Again it may be thought to introduce some bias as any faults or difficulties are likely to be duplicated over the two components. However it is in the nature of linear schemes that all candidates sit both components, and in modular schemes that effort is spread over several sessions, and thus at any one sitting there will be a much greater candidature than will be certificated, and indeed a greater mixture of abilities. The aim is to try to identify those questions which are apparently more difficult that others, in both schemes, and consider whether there is a greater likelihood of such questions appearing in one or other of the two examination regimes.

Syllabus Content

For this section of the research, the modules considered have been restricted to the first three pure mathematics modules, the first two mechanics modules and the first two statistics modules. A core of five (PM1, PM2, PM3, M1 and S1) which were in both the most common combinations for the single subject are taken as 'compulsory' with either S2 or M2 as the element of choice.

All syllabuses set out their aims and assessment objectives, and for the two syllabuses under the spotlight these are essentially the same. However there is in the modular scheme a set of objectives for each module, underlining the principle that each module must be regarded in syllabus terms as a separate entity. Clearly though, on the successful completion of either course, a candidate from either scheme should have developed broadly equivalent mathematical knowledge and understanding.

The syllabuses for both schemes are essentially very similar in content. Indeed if it were possible to include all modules in this analysis, the linear syllabus would be completely covered. However some differences do exist and it would be impossible with any legitimate combination of the six modules to duplicate the domain of knowledge within which aptitude had to be demonstrated.

Specifically, the linear scheme content is divided into two sections (see Appendix D). The first, entitled 'Pure Mathematics', contains 29 separate topics. Of these, all are covered by the first three compulsory modules except

(i) the remainder theorem (found in PM5)

(ii) the sum of an infinite series (PM4)

(iii) curves and equations in polar co-ordinates (PM4)

(iv) parametric expression of a point on a curve (PM5)

(v) matrices (PM4)

(vi) proofs, contradiction and induction (PM4)

The second section, entitled "Applied Mathematics' and encompassing both mechanics and statistics, is divided into two parts, the core syllabus and extension topics (which only appear in the question choice section of paper 2). All of the core is covered by the early mechanics and statistics modules except error analysis. This, together with numerical methods and differential equations of the extension, are found in the numerical analysis module (module 21) which in 1994 was wholly internally assessed and could not be used for comparative purposes.

Again most of the other extension topics are covered by S1, S2, M1 or M2 except

(vii) Hooke's law and SHM (found in M3)

(viii) random variables, probability distributions (S3)

(ix) sampling distribution, estimators, confidence limits (S3)

(x) hypothesis and significance tests (S3)

The topics included in the specified modules (see Appendix C), but not in the linear scheme, can be itemised as follows

(i) numerical integration (PM1) only appears as an extension topic in the linear syllabus

(ii) complex numbers (PM2)

(iii) correlation coefficients and simple regression analysis (S2)

It would appear that there are more topics left uncovered by the most common modular choices than by the equivalent linear scheme. However, in mitigation, the amount of choice available in the linear scheme would allow a number of topics to be omitted. Because there is no available evidence (even analysing the choice of questions only indicates which have been attempted not those which can be) it has to be assumed that all items are covered. It is thus inescapable that a candidate studying for the OCSEB A level linear syllabus will probably have covered more topics (or some topics in more detail) than the average modular candidate. However the evidence of attainment, and hence grade, is solely on the basis of questions answered, not what could have been, and the amount of evidence is likely to be very similar, regardless of origin and the equivalence of grading standards will not necessarily be compromised by these findings.

Another consideration is the differing coursework requirements of the two schemes. For the linear scheme it is a separate add-on component. For the modules which contain a coursework element it is a fundamental part of the assessment and indeed it was, in part, the idea of coursework being an integral part of the assessment which led to the original development of the MEI structured mathematics scheme.

In the linear syllabus, coursework consists of two pieces of work with the objective of assessing candidates' ability to apply their knowledge and skills in a meaningful

way and to present a coherent report on their work. Although tasks are set by the board, schools are encouraged to develop their own ideas for coursework.

The emphasis is somewhat different in those modules which contain an internally assessed component. In each case the module syllabus content is specified, with a particular element nominated as the subject for a coureswork study. For PM3, the numerical methods section of the syllabus is entirely internally assessed.

For M1 there are two coursework assignments, one on modeling and one experimental, which both form an integral part of assessing achievement of the stated module objective of introducing students to 'the concept of mathematical modeling and the processes involved'. The M2 objective is to instill understanding of the basic concepts of mechanics and the experiment which is a part of this aim is a coursework assignment.

The S1 module concerns itself mainly with understanding data and simple probabilities with the coursework involving data handling. The coursework assignment for S2 is specifically on regression analysis which, together with correlation coefficients, are not included in the linear syllabus.

The Question Papers

The structure of the question papers within each scheme, are, as may be expected, very different. For each component in the linear scheme the question paper is divided into two parts. Part 1 consists of compulsory questions (though the number is different for each component) and part 2 contains a number of questions (again different for each component) of which three must be chosen. These three together attract 60% of the marks on the paper and are the highest scoring at 30 marks each. The assessment scheme for the syllabus is completed by a 20% coursework element.

In contrast, each module paper contains four or five questions, each of which is compulsory. Of the seven modules considered, five also have a coursework element and for six modules, the total coursework would usually be about 20%. Coursework has, of necessity, to be excluded from consideration in this part of the analysis due to lack of data.

Although the fundamental differences which underpin the two types of assessment should not be underestimated, it is worth noting a number of similarities which existed in 1994. Firstly, the total time for the written papers was the same in each scheme. The two linear papers were each of three hours duration, the six module papers an hour each, so that to achieve an A level each candidate would be expected to spend the same amount of time answering questions. The coursework element was, in total, very similar for both schemes. Much of the subject content was the same and it could be expected that the questions would sample a similar proportion of this content. Table 7.1 outlines the composition of each paper together with the total number of marks and the percentage each paper contributes to the total assessment - which for the linear scheme would be at syllabus level, for the modular scheme at module level. (In line with the philosophies of modular and linear schemes, linear components are weighted as part of a whole, whereas modules are considered entities in their own right and the written paper weightings are given as a percentage of an individual module. Each module contributes 16.7% to the overall assessment.)

Table 7.1: Question Paper Format

	NUMBER OF	QUESTIONS	TOTAL	WEIGHTING
	Compulsory	Optional		
Component 1	8	3/5	150	40%
Component 2	6	3/8	150	40%
Module PM1	5		60	100%
Module PM2	5		60	100%
Module PM3	4		50	83.3%
Module M1	4		40	66.7%
Module M2	4		50	83.3%
Module S1	4		50	83.3%
Module S2	4		50	83.3%

Whilst the syllabus content in terms of knowledge of mathematical principles and processes is similar, the coverage of this domain will be different and, it is argued, more superficial in the case of modular examinations.

It is notoriously difficult to assess question demand. Question/paper setters use their own experience and judgement to specify how many marks a particular question is worth, especially in those papers where there are a variety of question

types which may all reflect different mark values. Wood (1991) believes that "this is not a topic susceptible to research" although what little has been done suggests that the more open-response questions actually attract more marks pro-rata than short answer questions. Wood presumes there is some time relationship i.e. a question attracting 40 marks takes about twice as long as a question attracting 20 marks.

Although it is not possible within the context of this research, it would certainly be possible to test the theory that question tariff is related to the time taken to answer a question. However, if all questions on a paper are compulsory there is less of a problem than when there is question choice (Willmott and Hall, 1975) which exacerbates the problem by giving equal tariff to optional questions, which later analysis shows to have been very unequal in demand.

Whilst recognising that defining question difficulty is far from straightforward, juxtaposing those data that are available i.e. the question setters' estimate of difficulty, defined by question tariff, and the time taken for the papers, one way of investigating question 'depth' is to assign to each question a time. The total for each scheme is 360 (six hours) and it can be assumed (from the lack of complaining feedback) that most of that time, in either scheme, was profitably employed by the candidates in answering questions. Detailed breakdown (see next section) indicates that few candidates were unable to at least attempt all the required questions, but it is in the nature of examining that some candidates are very much slower than others.

It may be assumed that the higher the question tariff, at least within each scheme, the more difficult or complicated the question and the longer it will take to answer. If we also assume that the time taken to answer a question is proportional to its tariff (within the context of the paper) then we can construct a common scale regardless of a question's origin. Thus a compulsory question on a component attracting 8 marks out of 150, say, will have a time value of 9.6 because a three hour paper has a time value of 180 equating to 150 marks, so 8/150 is equivalent to 9.6/180.

Additionally, each of the papers can be split up into areas of knowledge based loosely on the categories found in the modular syllabuses (see Appendix C). For Pure Mathematics these would be:

Algebra, Geometry and Trigonometry, Calculus, Complex Numbers, Vectors

Mechanics questions may cover:

Force, Newton's Laws of Motion, Equations of Motion, Centre of Mass, Work and Energy, Momentum and Impulse

For statistics the list contains:

Data Handling, Probability, Permutations and Combinations, Population Distributions, Hypothesis Testing

However, these categories have proved too wide too attempt to describe syllabus coverage in any detail. One could find differentiation of kx^n in the same calculus category as the solution of differential equations. Many additional skills are required in the latter case, even allowing that knowledge of the former would be required.

Therefore it was decided to categorise each question (or sub-question) according to the categories set out in Appendix D taken from the linear syllabus. These are in three parts. There are 29 categories in the pure mathematics section. These will be referred to by numbers. Part 1 of the applied mathematics syllabus (which encompasses both Mechanics and Statistics), the core, will be labeled category AM1 to AM10. Part 2, the extension topics, will be labeled NA1-NA3, ME1-ME6 and P1-P6. These are far more detailed, but do, in some cases, prove difficult to apply in the simple manner required. There are two reasons for this:

(a) many categories are subsets of others. Thus knowledge of category 11 (trigonometric functions) is required for category 12 which may be required for category 13 and so on. In these cases, the question has been classified once by the category which requires the highest order skills. In the final analysis some categories may not appear explicitly because they have been subsumed by another. This contrasts with those lower order levels which might appear only in one context and imply somewhat simpler questions have been asked.

(b) some questions are multi-faceted in that they appear to involve two or more different and apparently unconnected categories. For example a proof by induction of a formula involving matrices. Clearly knowledge of both categories which embrace these skills is required. In some cases it is possible to determine the major element as, for example in an integration question for which a sketch is required. Such a question would be classified as, say category 25 (which may involve

category 24 as a subset) and not category 8. However, when two unrelated categories appear indivisible within a question, the marks are shared evenly between them.

Not all decisions are clear cut, though many are, and where there is apparent neglect of a given category, apart from those identified above, many can be identified as being united in some way with another class.

The total scores available in each category enable a direct comparison of the constituent parts of each of the papers. However, it should be noted that although all these topics are represented in the syllabuses, they do not necessarily all have to appear within a session's question papers. Thus even if a topic is not represented in the papers for the summer of 1994, this does not mean the topic is never examined within that syllabus.

The table below outlines the results of the categorical analysis for each paper with two or more numbers signifying that more than one question on the paper was based on the category given (the category number refers to those of Appendix D). A time value is assigned to each question (instead of the raw marks) so that the category coverage can be judged according to a common scale. The conversion factor from raw to weighted for the linear scheme is 1.2, for PM1-PM3 the factor is 1, for M1 1.5 and for M2, S1 and S2 again 1.2. Aggregating over each syllabus the total coverage in each category can be gauged.

Table 7.2: Syllabus Coverage for All Written Papers (Weighted Marks)

CAT	C1	C2	PM1	PM2	PM3	M1	M2*	S1	S2*
1	-	13.2	5	-	3	-	-	-	-
2	4.8	-	-	-	-	-	-	-	-
3	-	-	3	-	-	-	-	-	-
4	9.6*	12	-	5	-	-	-	-	-
5	4.8, 36*	-	12	12	-	-	-	-	-
6	-	10.8	2	-	-	-	-	-	-
7	36*	12	-		4	-	-	-	-
8	6	10.8	2,2		-	-	-	-	-
9	7.2*	13.2	7		-	-	-	-	-
10	3.6	-	-		-	-	-	-	-
11	-	-	12	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-	-
13	-	-	-	-	-	-	-	-	-
14	-	-	-	2	-	-	-	-	-
15	22.8*	-	-	6	-	-	-	-	-
16	13.2*	-	-	6,3,2	-	-	-	-	-
17	-	-	-	4	-	-	-	-	-
18	8.4	-	-	-	12	-	-	-	-
19	8.4	-	-	-	4	-	-	-	-
20	8.4, 4.8*	-	-	-	-	-	-	-	-
21	9.6, 15.6*	-	-	6	3	-	-	-	-
22	-	7.2*	-	-	3	-	-	-	-
23	-	-	7	2	3	-	-	-	-
24	-	-	-	3	-	-	-	-	-
25	8.4, 9.6*	-	-	9	-	-	-	-	-
26	-	-	-	-	-	-	-	-	-
27	6*	-	8	-	-	-	-	-	-
28	6, 19.2*	-	-	-	-	-	-	-	-
29	3.6	-	-	-	-	-	-	-	-
AM1	-	36*	-	-	5	-	-	-	-
AM2	-	15.6*	-	-	-	-	-	-	-
AM3	-	-	-	-	-	-	-	-	-
AM4	-	-	-	-	-	-	-	-	-
AM5	-	-	-	-	-	7.5	-	-	-
AM6	-	-	-	-	-	-	-	-	-
AM7	-	-	-	-	-	-	-	-	-
AM8	-	10.8*	-	-	-	-	-	14.4	-
AM9	-	22.8*	-	-	-	-	-	14.4,15.6	-
AM10	-	-	-	-	-	-	-	-	-
NA1	-	18*	-	-	-	-	-	-	-
NA2	-	10.8*	-	-	-	-	-	-	-
NA3	-	-	-	-	-	-	-	-	-

ME1	-	13.2*	-	-	-	6,7.5	14.4,6 ,7.2	-	-
ME2	-	7.2*, 21.6* 9.6*	-	-	-	9	9.6	-	-
ME3	-	14.4*	-	-	-	15	8.4,6	-	-
ME4	-	13.2*	-	-	-	-	8.4	-	-
ME5	-	-	-	-	-	-	-	-	-
ME6	-	13.2*	-	-	-	15	-	-	-
P1	-	13.2*	-	-	-	-	-	-	15.6,1 4.4
P2	-	-	-	-	-	-	-	-	-
P3	-	36*	-	-	-	-	-	-	14.4
P4	-	14.4*	-	-	-	-	-	-	-
P5	-	-	-	-	-	-	-	-	-
P6	-	10.8*	-	-	-	-	-	15.6	-

* denotes choice

Additionally 13 marks were available on PM3 for a question on complex numbers and 13 marks on S2 for a question on correlation coefficients.

Firstly there are a number of categories not covered in the examination for either syllabus. Categories 12 and 13 were not found explicitly, but knowledge of angles between lines and planes was certainly used in mechanics questions. Category 26 was subsumed in 27. AM3 was covered in some way by coursework. AM4 and AM10 just did not appear (though again they were possibly included in some coursework). AM6 and AM7 were subsumed by the slightly wider categories of ME1 and ME2. Euler's method did not appear on the written papers, neither did Hooke's law, the binomial distribution or sampling statistics.

For the rest, the differences between the two schemes are summarised by adding together all those contributions from the linear and modular scheme as detailed in table 7.2 and subtracting the modular coverage from that of the linear scheme. For example, in category 1 of table 7.2, the weighted marks shown for linear questions is 13.2. For the same category there are two contributions from modular question papers of 5 and 3. Subtracting 8 from 13.2 gives 5.2 as shown below:

Table 7.3: Difference in Syllabus Coverage

CAT	LIN-MOD	CAT	LIN-MOD	CAT	LIN-MOD
1	5.2	17	4.4	AM1	-31.0
2	4.8	18	-3.6	AM2	15.6
3	-3.0	19	4.4	AM5	-7.5
4	16.6	20	13.2	AM8	-3.6
5	16.8	21	16.2	AM9	-7.2
6	8.8	22	4.2	ME1	-27.9
7	44	23	-12.0	ME2	19.8
8	12.8	24	-3.0	ME3	-15.0
9	13.4	25	9.2	ME4	4.8
10	3.6	27	-2.0	ME6	-1.8
11	-12.0	28	25.2	P1	-16.8
14	-2.0	29	3.6	P3	21.6
15	16.8	NA1	18.0	P4	14.4
16	2.1	NA2	10.8	P6	-4.8

Although it appears from the table above that syllabus coverage is much greater in the linear scheme, it is important to recognise that, of the values assigned, 252 are redundant in the linear scheme and only 60 in the modular scheme because of the nature of the choices involved.

If the table is re-calculated including only those questions/modules which are compulsory, there is a certain amount of change.

Table 7.4: Comparison of Compulsory Questions

CAT	LIN-MOD	CAT	LIN-MOD	CAT	LIN-MOD
1	5.2	17	-4.0	AM1	-5.0
2	4.8	18	-3.6	AM2	-
3	-3.0	19	4.4	AM5	-7.5
4	7.4	20	8.4	AM8	-14.4
5	-19.2	21	0.6	AM9	-30.0
6	8.8	22	-3.0	ME1	-13.5
7	8	23	-12.0	ME2	-9.0
8	12.8	24	-3.0	ME3	-15.0
9	6.2	25	-0.6	ME4	-
10	3.6	27	-8.0	ME6	-15.0
11	-12.0	28	6.0	P1	-
14	-2.0	29	3.6	P3	-
15	-6.0	NA1	-	P4	-
16	-11.0	NA2	-	P6	-15.6

As was indicated previously and becomes more obvious when the optional element is eliminated, the syllabus coverage in the linear scheme (in comparison with the modular) is greater in the more traditional areas of pure mathematics concerned with more complex functions and algebra and detailed mathematical relationships. The modular scheme presents candidates with a greater variety of challenges though probably losing out in some of the depth of coverage.

One of the by-products of this analysis was a somewhat greater appreciation of the differences in the two types of question paper. Certainly the modular questions were far more structured (i.e. the first part of each question would be accessible to the majority of candidates, the last part to very few) and for this reason alone may have seemed easier. The two approaches can be illustrated by way of a simple example. If a candidate were asked to find the turning points of some defined polynomial function f(x), the traditional scheme would ask just that. The modular approach would be to ask the candidate to differentiate the function (a), factorise the differential (b) and finally state the value of the turning points (c). The additional accessibility is clear. Its advantage is that good candidates can demonstrate their skills equally in either scheme, but weaker candidates can at least attempt the first part with some confidence (differentiation, for example, is usually a very well learnt skill even by the less able).

The structuring also had the effect of making the modular papers less daunting. It appeared that more effort had been put into their "user friendliness" . The effect is difficult to quantify, though it would be possible to ask the same question in a number of different ways and measure the responses for a similarly representative samples of students.

Also in a number of questions in the modular papers candidates were asked to comment on their answers. A certain amount of thought and understanding is required for this. The correct application of a learnt technique to a recognisable situation does not always imply understanding of what is being expected in the wider context, a less recognisable situation say, that is an objective assessment of the mathematical requirements needed to solve a given type of problem. There is some emphasis on this type of understanding in the modular scheme.

It is clear that differences do exist, depth has been exchanged for a broadening in the skills and knowledge requirements of the syllabus. It is unlikely that either could be said to be more demanding than the other, but they are different.

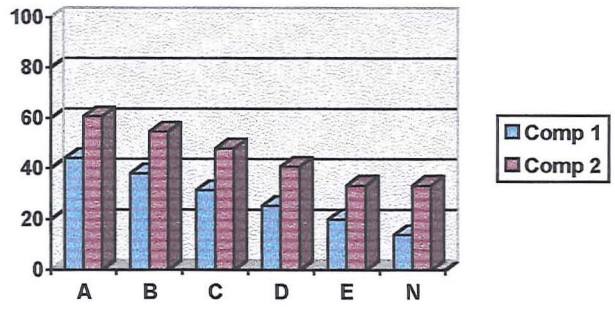
Finally in this section, before investigating performances on the individual questions, the overall 'standard' of each paper must be considered. One method, probably the only method, is to examine the percentage of the total marks awarded at each grade. This, of course, begs the question of comparability of standards not only between modular and linear schemes, but also of standards within the two schemes. The previous chapter served to show that grade standards within the modular syllabus were consistent from session to session and in their relationship with each other, and there is no reason to doubt comparability within the linear scheme.

The drawing of grade boundries is fundamental to the setting of standards in an A level examination. It is the weakness of any comparability study in that there is no direct measure by which grades can be judged and it is almost impossible to gauge with any accuracy the standard of an individual module as compared with that of an individual component, not least because comparisons of this type are usually only made at syllabus level (although in 1996 there was an attempt at a comparison of the boundary standards of components and modules). It is acknowledged that the remainder of this section rests on the unsupported assumption that grade standards between modular and linear schemes are reasonably consistent. This assumption rests on the judgemental competence of the awarders, some of whom are common to both linear and modular schemes, and all of whom have knowledge of linear awarding procedures. However there is none of the statistical backup between schemes which normally attends awarding meetings in order to ensure year on year comparability because none exists. Figures 7.1, 7.2 and 7.3 are therefore presented to illustrate the percentages required to attain the various grades within each scheme, with the proviso that the comparisons which are drawn between modules and components are only valid if the assumption of equivalence of grading standards holds. This is the same assumption which has been made throughout this research and one that is based on the notion of social comparability.

The figures below indicate the percentage of marks required for the various grades, and should the assumption of comparability hold, could also be used as indicators

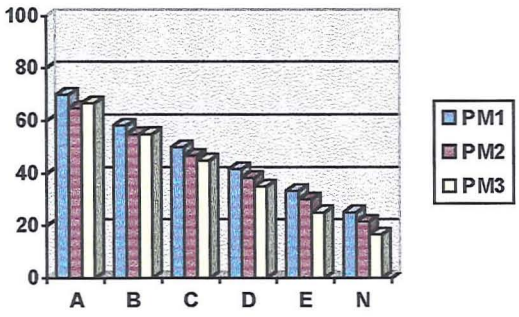
of the relative difficulty of the components/modules i.e. the fewer marks required to obtain an A, say, the harder the paper. (It should be noted that since 1994 all components are graded separately and aggregated to give a syllabus boundary).

Figure 7.1: Percentage of Component Marks Required for Each Grade



The chart in figure 7.1 above indicates that, in the opinion of the awarders, component 2 was somewhat easier than component 1, but that, on average, about 52% of the marks would be required for an A, and about 46% for a B on the written papers. It may be worth noting that this differential may be more pragmatic than real. If awarders had needed to make adjustments to grade boundaries "in the light of statistical and technical evidence" in order to ensure year-on-year comparability, they may have chosen to make any adjustment on one component only.

Figure 7.2: Percentage of Module Marks Required for Each Grade - Compulsory Modules



For module PM3 (and all statistics and mechanics modules considered in figure 7.3), the percentage of marks required at each grade includes a proportion of coursework marks. Boundaries for the coursework and written elements of each module were not set separately until January 1995. It appears from figure 7.2 above that awarders considered PM3 to be the most difficult, hence the lower grade

boundaries, except at A where apparently module PM2 provided the greater challenge.

Figure 7.3: Percentage of Module Marks Required for Each Grade - Optional Modules

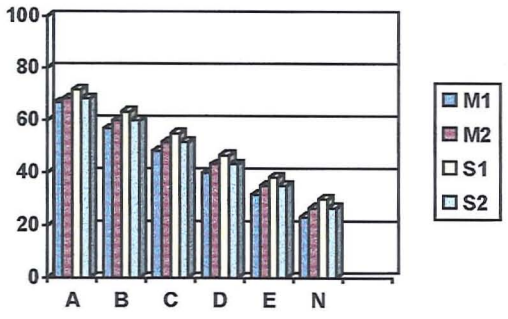


Figure 7.3 suggests that the mechanics modules were probably more difficult than the statistics modules, and that the second mechanics module was more accessible than the first.

Bearing in mind the separate awarding processes, and the nature of that process, there is some similarity between the two schemes. The evenness and pattern of the grade mark bandwidths are sufficient to demonstrate this although, in the opinion of the awarders the linear scheme is much harder. If the assumption of comparability of grade standards holds then it appears that in standard of paper, linear candidates can afford to make many more mistakes than can modular candidates since comparable grades require over 10% less of the marks to be scored. This alone would tend to offset any perception of greater question demand.

Table 7.5 Percentages of Marks Required at Each Grade

	A	B	C	D	E	N
Linear	60.5	53.9	46.8	39.7	32.9	26.1
Modular	68.1	58.1	49.4	40.8	32.2	23.9

Table 7.5 gives the comparisons for percentage of marks scored at subject level, and this includes marks for coursework. (The modular scheme boundary is

synthesised by the aggregate of the raw mark boundaries for combinations of modules 1, 2, 3, 7, 13 and 8/14 in June 1994).

The inclusion of coursework narrows the difference in requirements for the attainment of a grade, although it is clear that the awarders expect less evidence at high grades, in terms of proportion of marks scored, in the linear scheme. At the bottom end of the grade range the expected performance is much the same regardless of scheme.

Question Performance

There are two indicators which can be used to judge quantitatively the demand of any question:

(I) the total marks awarded for the question

(ii) candidate performance on that question.

The total number of marks given to each question is a measure of its relative difficulty within the paper, at least in the eyes of the examiner. They are set in the expectation that few candidates, if any, will score all of them. Indeed the purpose of the question is to discriminate between candidates, and any which fails to do so is, in one sense, wasted. It is important that the scale of question difficulty as exemplified by the marks awarded for each question is in line with the scale of difficulty as experienced by the candidates. There are a number of reasons for this, including the fact that it is intuitive to give the highest level of achievement the greatest reward.

If easy questions gained high marks, and difficult questions low marks, then the mark range within which most candidates' scores fell would be very short. Thus discrimination would take place within a small number of marks. Reliability would become an even greater issue, and rank ordering would be very difficult.

A trivial example illustrates the problem. Suppose there are 10 candidates, all of whom can do the easy question, but are spread in their abilities (from completely unable to perfect) when faced with a more difficult problem. If there are 2 marks for

the easy question and 10 for the difficult then all ten candidates can be spread throughout the 2 to 12 mark range and can be properly rank ordered if abilities warrant it. If the questions were reversed in value then all 10 candidates would have to be fitted into a 2 mark range with a resulting loss of rank ordering.

The possibility of equating difficulty with discrimination within certain limits exists. However, a question that is too easy will be unable to discriminate (because most candidates will score high marks) but equally a question that is very hard will also lose discriminatory powers because few if any marks will be scored.

The quality of the candidates on whom this analysis is based can be judged by comparing the means for each paper with the grade that such a mark percentage would achieve. On component 1 the average mean mark indicates that the candidates are nearly B standard, on component 2 half-way up the D mark scale. Since, in this case, the same candidates are involved it indicates that there was a difference in grade standard between the two components with component 2 somewhat harder than the grading decisions would indicate.

For the three pure mathematics modules, the candidates are, on average, nearly B on the first module, still a C on the second, but of D standard on the third.

For both mechanics modules the candidates are, on average, of C standard, those for the second module possibly slightly better than those on the first. The same C standard is found for those taking the first statistics module, but the second paper candidates were again of D standard.

The apparent discrepancies do reflect the grade distributions for the papers, i.e. more candidates get high grades on, say, PM1 than PM3. Since the pattern of ability bands is similar in both schemes, as compared with grading standards, it is also reasonable to assume that the candidates are representative of their cohorts.

The figures in Appendix E illustrate the mark/frequency distributions of the questions on each of the papers. (A normal curve based on the same mean and standard deviation is super-imposed for reference). It should be remembered that one aspect of having a prescriptive mark scheme is that some combinations of marks are either impossible or almost impossible to achieve. For example on

question 3 of the first component, the mark scheme makes it impossible for any candidate to achieve a score of 4. Thus these distributions must be viewed with an eye to the possible.

What is immediately clear is the diversity of scoring within each of the questions for both types of scheme. However one observation that can be made is that the questions that did not work i.e. did not discriminate well between candidates, tended, in the linear case especially component 1, to be too hard. In the modular case it was because they were sometimes too easy. In either case they could have been omitted from the papers with little loss of paper effectiveness.

Another interpretation of the same result would indicate the existence of a ceiling/floor effect. In the linear case, the inability of a number of candidates to do some questions, e.g. question 7 on component 1, would imply that at the bottom end of the ability range, the assessment would be unable to differentiate between candidates and thus there would be a 'floor' effect i.e. a number of candidates on the same mark because of the deficiency of the evaluation process. The reverse, or 'ceiling' effect, e.g. question 2 of PM1, was observable in some modules where the questions proved too accessible for a number of candidates and failed to differentiate between them.

This leads to the observation that, in all probability, the perceived leniency or difficulty of an examination is based on the performance of candidates at the extremes of the ability range. The linear scheme may be seen as hard, the modular easy, though both examinations may discriminate equally well for the majority of candidates. What to some observers is the stretching of the most able, is to others a floor effect, and vice versa. The fundamental difference is that in the two component linear scheme, there is less chance to rectify the problem than there is in a six module assessment regime.

In a similar vein, though in many cases the distribution was relatively centrally located, such skewness as existed in the linear case tended to be towards the bottom end of the mark range. The reverse happened with the modular scheme. This is entirely consistent with the result that might be expected from the structuring policy described above.

Again the total mark distributions for the components have a mean value below the median of the mark range. The reverse is found for modular totals. However the differences are sufficiently small only to provide grounds for comment rather than any for criticism.

Discrimination and Facility

There are a number of standard techniques and parameters which are routinely used to demonstrate the effectiveness of a question within a paper. A question is usually classified by its facility value (i.e. its mean value as a proportion of the total marks available) and its discrimination coefficient. One measure of this is the question/total correlation coefficient shown below.

Table 7.6: Correlation Coefficients for Components

Question	Component 1	Component 2
1	0.55	0.63
2	0.67	0.45
3	0.49	0.58
4	0.62	0.67
5	0.55	0.43
6	0.50	0.65
7	0.52	0.73
8	0.74	0.81
9	0.73	0.80
10	0.74	0.46
11	0.83	0.84
12	0.66	0.72
13	0.77	0.77
14	-	0.72

Table 7.7: Correlation Coefficients for Modules

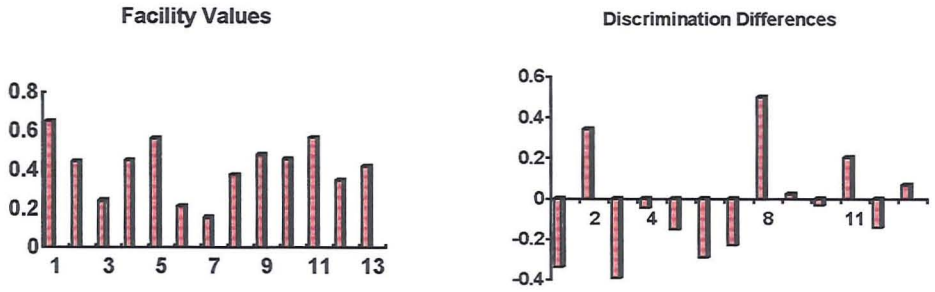
Qun.	PM1	PM2	PM3	M1	M2	S1	S2
1	0.79	0.72	0.80	0.85	0.81	0.64	0.73
2	0.52	0.81	0.85	0.82	0.77	0.75	0.84
3	0.80	0.75	0.83	0.85	0.71	0.78	0.87
4	0.70	0.68	0.80	0.76	0.78	0.83	0.80
5	0.65	0.72	-	-	-	-	-

There is a stronger correlation between the module items and the totals than found for linear questions, but this is probably affected by question choice and the somewhat narrower focusing of the module test subject matter as well as the reduced number of questions on the module papers.

A more sophisticated means of describing discrimination is again to use the covariance/variance of part with whole which has the property of being additive (Fowles, 1974) This latter measure is analogous to the module weighting used in the previous chapter since the ability of a question to discriminate effectively, i.e. its 'scatterability', is directly equivalent, numerically, to its weighting within the paper.

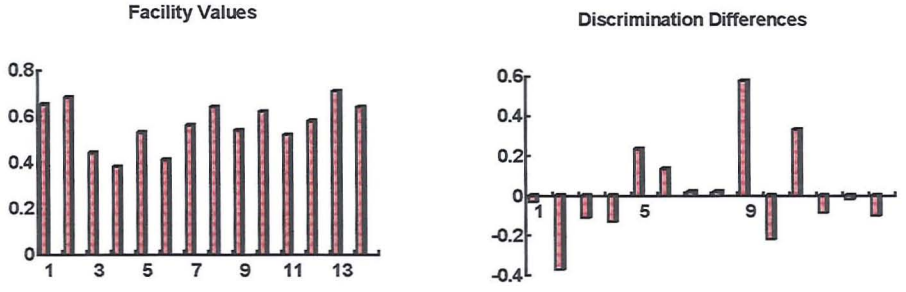
In figures 7.4 to 7.12 below illustrating facility and discrimination values for each component/module, each intended weighting (defined as the proportion of the total marks available for the question) has been subtracted from the achieved weighting and this difference has been scaled by dividing it by the intended weighting. Thus comparisons can be made directly because each difference is given as a proportion. Negative values indicate that the question is failing to discriminate as effectively as it should, given the total question mark.

Figure 7.4 : Facility and Discrimination Differences for Component 1



A number of features are observable from figure 7.4 which relate to question performance on component 1. Firstly there is no unambiguous connection between the ability of a question to discriminate effectively and its facility value. However, with the exception of question 1, there is a tendency of questions which are low discriminators to also have low facility values. This reinforces the point made previously.

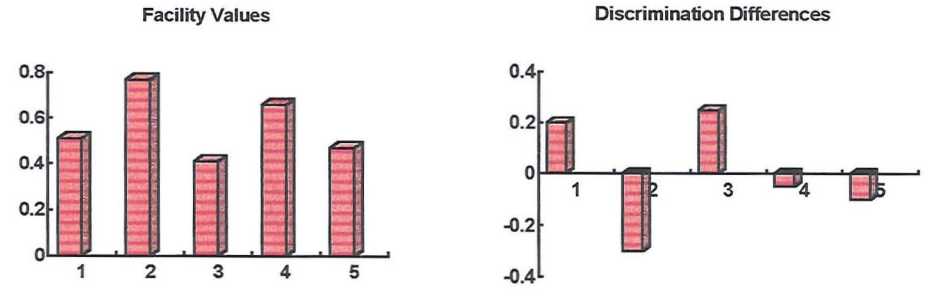
Figure 7.5 : Facility and Discrimination Differences for Component 2



Again in figure 7.5 based on question level data for component 2 there is no clear pattern emerging between the ability of a question to discriminate and its facility value, although the three 'best' discriminators are questions 5, 9 and 11 which are the three questions with facility values closest to 0.5. This is neither an unusual or unknown finding, but it is not seen in component 1 questions and emphasises that differences between components should not be underestimated. The best discriminator was only taken by 40% of the candidature so its effect would not have been universal.

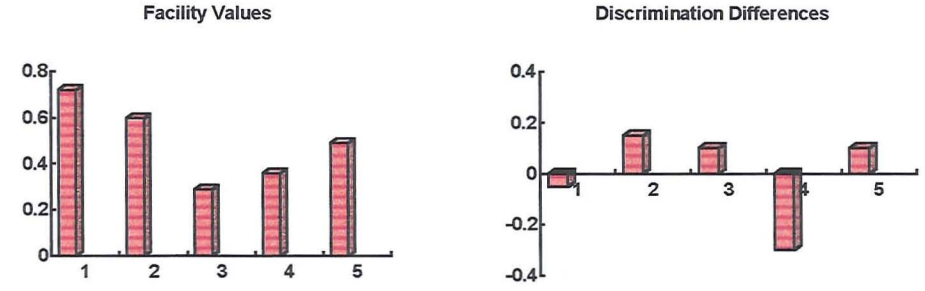
The modular picture, illustrated in figures 7.6 - 7.12, also shows some interesting differences in the performance of the questions.

Figure 7.6 : Facility and Discrimination Differences for Module 1



In figure 7.6 which illustrates module 1 performances, although question 1 (facility value 0.51) does discriminate well, it is not as good as question 3. In contrast to component 1, but in a similar manner to component 2, the question carrying the lowest weight is one which attracts the highest scores. It seems to be a feature of this particular modular scheme - that poorly discriminating questions are ones which are the easiest to answer.

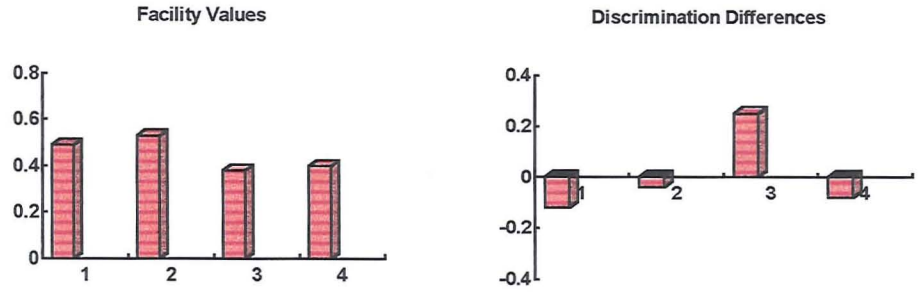
Figure 7.7 : Facility and Discrimination Differences for Module 2



In figure 7.7 above which relates to module 2 data, question 4 stands out as a poorly discriminating question. It has a low facility value (though not the lowest on this paper) and would appear to be one of those questions where the first part was fairly easy, but the last sufficiently difficult to cause very low scoring. Thus this module was somewhat unbalanced, a feature also of the previous module. Although such lack of discrimination is by no means confined to the modular scheme, its effect is somewhat more profound because the number of questions asked is so much fewer than in the more traditional component. Also of note in this

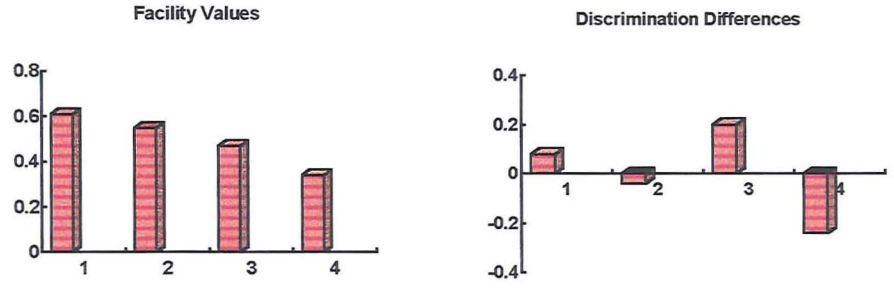
module is the low facility value of question 3 - a more unusual finding in the modular context.

Figure 7.8 : Facility and Discrimination Differences for Module 3



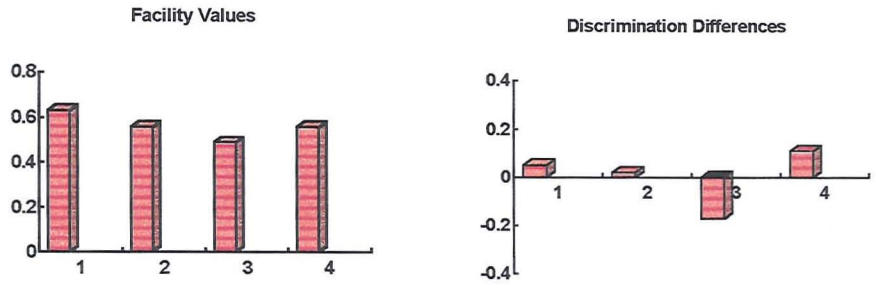
Here, in figure 7.8, module 3, we see that the lowest scoring question (though only just) is the most highly discriminating, and it is difficult to discern any consistent pattern establishing itself. Relatively equivalent facility values indicate a very even question performance across the paper.

Figure 7.9 : Facility and Discrimination Differences for Module 7 (M1)



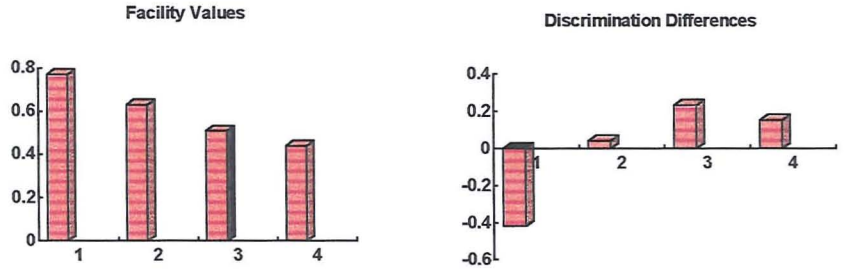
In contrast to the previous module, the lowest scoring question on module 7, illustrated above in figure 7.9, is also the poorest at discriminating between candidates and it is obvious that question 4 is the least effective question on the paper, probably because it was too hard for most candidates.

Figure 7.10: Facility and Discrimination Differences for Module 8 (M2)



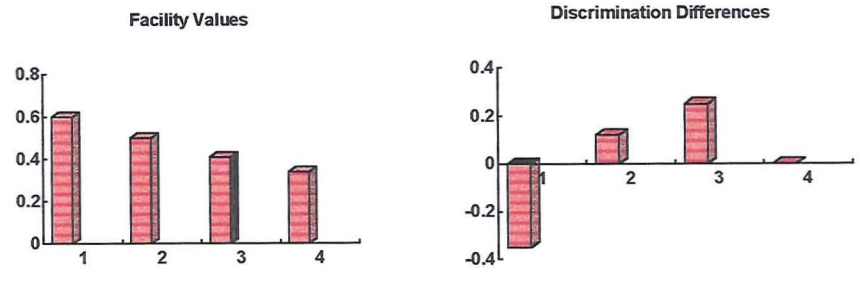
The indications from figure 7.10 above is that the high scoring module 8 discriminated rather well on all questions indicating a balanced paper which targeted candidates accurately.

Figure 7.11 : Facility and Discrimination Differences for Module 13 (S1)



A very high scoring question on module 13 (see figure 7.11) has led to loss of discrimination (because it is too easy) and the differentiation between candidates is made mainly on the other three questions. The paper seems therefore somewhat unbalanced.

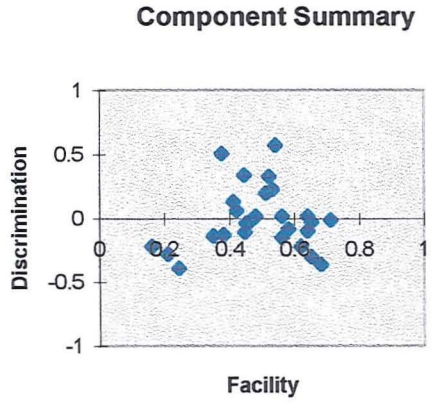
Figure 7.12 : Facility and Discrimination Differences for Module 14 (S2)

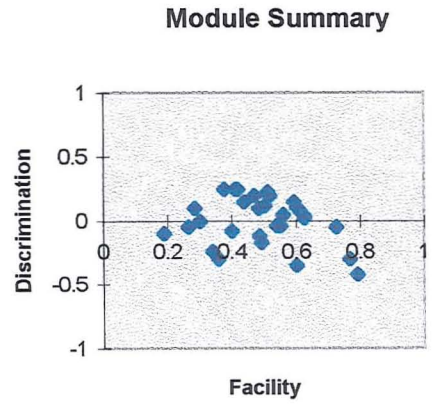


Module 14 exhibits in figure 7.12 a very similar pattern to the other statistics module with one poor question. Whilst this is unlikely to be an artifact of statistics per se it is an odd finding.

Although individual module patterns are of interest, of more immediate concern is the contrast between components and modules. A summary of the data is given in figure 7.13 and table 7.8 below averaged for component compulsory questions, optional questions and for modules.

Figure 7.13: Summary of Facility and Discrimination Values for Components and Modules





As is self-evident from the summary graphs above, there is far more variation in the component questions regarding discrimination indices, whereas facility values extend slightly further up the scale for modular questions. These are also reflected in the table below:

Table 7.8: Summary of Component and Module Data at Question Level

Type	Correlation Mean	Facility Mean
Component Compulsory	0.575	0.44
Component Optional	0.74	0.54
Module	0.77	0.49

N.B. There is little point in giving a discrimination mean since positives and negatives will cancel out to leave a zero, or near zero, result.

As a general rule it would appear that there is more similarity between the module questions and the optional questions of the components in that both carry higher correlations (not surprising in view of the higher proportions of marks contributed to the total) and also have higher facilities, which is not quite so expected. However there are differences between components as highlighted in the detailed analysis and in this context, it is clear that component 1 not only contains a number of very difficult questions which must, of necessity discriminate badly, but in order to compensate has three very highly discriminating questions. In general, the linear scheme would appear to have a greater diversity not only in facility values (especially for component 1) but also show a greater variance in the ability of questions to discriminate. This is partly an artifact of a greater number of questions which almost inevitably will show differences in effectiveness and, at least in part, a result of question choice. (It is possible to make a modification to the facility values (Morrison in Nuttall and Willmott, 1972) of those questions which are optional in order to take account of the ability of the group making the choice. The adjustments are typically O(0.02) and are of little consequence in this analysis). Arguably it is more important in a limited question environment for questions to be more effective both in accessibility and discrimination, and this is generally true for the modules examined here. However there is no clear connection between facility and discrimination values; in particular there appears to be no pattern which is recognisably either modular or linear.

The subject matter which attracts facility values at the extremes of the range is very variable. From the analysis above, the questions which are found hardest, irrespective of context, are those which involve vectors. Conversely, data handling seems to be sufficiently straightforward to attract high facility values, again irrespective of context. Other subject areas are less easily categorised. A possibility therefore must be that the complexity of the questions can, for some mathematics subject areas, more easily disguise the difficulty of the mathematics concepts involved. One example might be a fairly simple mechanics problem which involves so much arithmetic that candidates make silly errors (especially where signs are concerned) and thus get low scores. There is little evidence from the results above that some topics are more accessible (or otherwise) within the modular context.

Reliability

Leaving aside the considerable question of examiner reliability which, for the purposes of this thesis, is assumed to be the same for each examination, (and this is a reasonable assumption for mathematics where examiner reliability is high) another important facet of examinations is that aspect of reliability usually known as internal consistency. Previously the modular syllabus reliability as demonstrated by the consistency across modules was discussed, but in this instance the comparison is between components and modules. Therefore the consistency of individual elements, as demonstrated by inter-item correlations within each component/module, needs to be calculated.

Two additional factors are of relevance; those of question choice and length of examination. The normal method of determining inter-item correlation is to use Cronbach's alpha as used previously. However, where there is question choice a modification to this method is used, namely Backhouse's P (Backhouse in Nuttall and Willmott, 1972). This reduces to alpha when no choice is available. Both alpha and P give lower bounds to the reliability estimate, but can be used to compare components and modules with a fair degree of rigour.

However, test reliability increases with its length. An adjustment, known as the Spearman-Brown formula is available to modify the reliability coefficients such that,

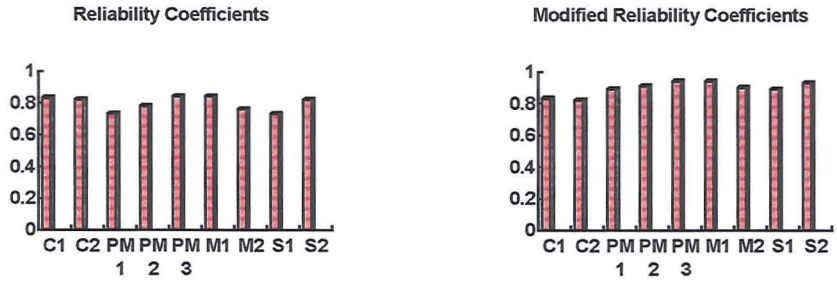
were the test longer, an estimate can be made of the consistency of the longer test based on the mean item correlations of the actual test. The formula is given below:

$$\alpha_n = \frac{n\alpha_1}{\{1 + (n - 1)\alpha_1\}}$$

n = any ratio of test length to that for which α_1 is known
α_1 = reliability of test of unit length
α_n = test of n times unit length

Using this transformation, it is possible to compare the reliability of modules directly with those of components by converting the reliability of a module of unit length to one three times as long.

Figure 7.14: Reliability Coefficients for Each Component/Module



Although there would appear to be little difference in reliability if no account is taken of the difference in the lengths of each component/module, once comparison is made with no time differential is fairly clear that modules are inherently more reliable than components. This is hardly a surprising finding given the more restricted nature of the questions within modules and the more heterogeneous nature of components, and with this rather specialised definition of reliability may be considered somewhat irrelevant, but it does nevertheless point to an advantage, albeit small and predictable, that modules possess over components. It is also a result one would expect in most modular examinations, irrespective of the nature of the subject. Direct equivalence to components would involve a further adjustment to allow for the internal consistency over a combination of three modules, but because there are few common candidates between these modules, the sampling errors of the statistics in the formula used (see Nuttall and Willmott, 1972) are too high for satisfactory results. However, the internal consistency between modules is high (see chapter 5) and it is unlikely that the reliability would be reduced much from the above estimate.

Populations and Gender

There has been very little emphasis placed on gender differences. The main reason is the overwhelming preponderance of males in the traditional scheme. Any results are therefore based on a small sample which is probably unrepresentative because many of the females will be from the, now co-educational, sixth forms of traditional male public schools. In the same way, centre type has not been considered because most of the linear candidates are from the independent sector.

Of the 378 candidates for the linear examination, 54 (or just over 14%) of the candidates were female, for the modular scheme 1315 out of 4369 (i.e. 30%) were female, more than doubling the ratio of females to males. Of these 378 candidates, over 99% came from the independent sector, generally high achievers in public examinations. This contrasts with the 36% for the modular scheme. These data illustrate that the populations for the two examinations are, at least superficially, very different and that the results from this section must be considered with this in mind. Additionally, since the components/modules from which the question paper analysis was derived were all from the summer 1994 examination session, there was potentially an even greater diversity within the module population caused by the usual mix of gender, age and status (i.e. first time takers, resitters, first year sixth, terminal candidates and so on).

Whilst these factors do not invalidate any of the conclusions, they might suggest that some further analysis might be done especially regarding gender, the distribution of the sample between centre types being both too stratified (linear) and too diverse (modular) to allow any rigorous conclusions. Despite the choosing of the sample being completely random and not stratified in any way the question papers were split in the proportions expected i.e 16% of the linear scheme's scripts were from females and about 30% of the modular scheme. However, regarding modules, the sample also exhibited to a rather inflated degree the choices of module. Although for the compulsory modules just under 30% were female, for the statistics modules this percentage increased by 10 to 40%, and for the second mechanics module dropped to under 20%.

Whilst for most questions the facility values were very similar for both males and females, some questions appeared to have facility values which were gender related. Question 11 on component 1 showed a difference in facilities of 0.17, with males outperforming females. The question (on algebraic inequalities) is not obviously biased in any way, and the observed differences could be as much a result of infelicitous choices on the part of the females as any inherent question bias. On component 2 very few differences are observed, and none of any consequence. However, question 3 on module 1 shows the same contrast in facilities, and again the topic of the question is an algebraic calculation. Conversely on module 2 females are gaining higher facility values on the trigonometry questions. In neither case can these differences be attributed to question choice since there is none. Of the other modules only one produces notable facility value differences. This is on the first statistics module where males outperform females on two questions involving probability and permutations and combinations. It is only possible to surmise that the simple algebra, which is at the heart of such calculations, may be one reason for the observed pattern, but that would be in line with other observations.

What is of more relevance here is whether there is a gender factor which can be identified as uniquely modular or linear. Even allowing for the sparse sample of females, there is little evidence to show an assessment effect at question level, although there may well be a link between the subject matter and performance.

Discussion

In the requirements of learning as laid down in each syllabus, there would appear to be some difference in the amount of mathematical skills required by each candidate in the two schemes. Thus the potential domain of assessment for linear schemes appears somewhat wider, although part of the requirement is optional.

However, in detailed consideration of the content of the question papers, it is fairly clear that syllabus coverage is greater in the modular scheme, in other words sampling of the domain is more effective. This should improve the validity of the scheme in the sense that the content validity (i.e. how adequately a specified universe of content is sampled (Ebel, 1965)) of a module is higher than that for a component. There is evidence that the linear scheme requires a deeper knowledge

of some topics to compensate for this lack of breadth. However, there is little evidence to suggest that the validity of the modular scheme is at risk through construct under-representation.

It is also clear that some of the linear questions are proving very difficult, especially for weaker candidates and that modular question structuring is enabling a more positive performance to be seen.

It is impossible to make a quantitative judgment as to whether one scheme is superior to the other. That they are different has been amply demonstrated. That they are both able to discriminate effectively has also been seen. What may, tentatively, be concluded is that even weak candidates can get something out of a modular scheme. They can demonstrate positive achievement. It is far more difficult in a linear scheme for the less able to do this. Questions are harder (as shown by the means found) without being more effective in terms of discrimination.

In general, questions from both schemes do discriminate fairly well, but not equally. However, there is no evidence that such differences are in any way illegitimate. What is more evident is that in a good testing environment, some component questions are too hard to discriminate well, and one or two modular questions too easy. It would however be a mistake to generalise too much on the basis of one or two untypical examples.

There is some evidence that topics which elicit poor responses do so whatever the environment. Vector questions are hard; data handling questions tend to be easy. There is also evidence that whilst there are gender differences in performance these are not confined to one scheme of assessment and again are probably based on the question subject matter rather than any difference in approach to the question setting.

It may be argued that the traditional approach to examining is a better preparation for university, or at least some universities, but this is difficult to sustain as more institutions of higher education turn to modular curricula. Even if there were some truth in the assertion, it is questionable whether this is the only purpose of an A level examination, or even the most important one. Increasingly A levels are seen as a school leaving examination, with all the purposes that implies. With the modular

scheme on offer it is always possible to choose to go on to study the more difficult topics if such preparation is what is needed. That is the strength of the modular approach.

What may be deduced from this analysis is that the two schemes have more in common than might at first be apparent. Harsh judgements may have been made from evidence that is, at best, untypical. The subject matter covered by the examination processes is broadly similar, as is question performance. Although facility values are in some instances lower in traditional schemes, it is arguable that these are poor questions and probably have no place in the overall scheme. A similar argument could be mounted regarding the very high facility values for some module questions, but these go far in explaining the differences in percentages of raw marks required to obtain the high grades between the two schemes.

In terms of comparability, of demands on the candidates in respect of the amount of learning and difficulty of questions, there is evidence that there is little cause for immediate concern. If we accept the premise implicit in the social definition of comparability that grading standards are the same for both schemes, then the picture presented here, of broadly equivalent demand but better performance because of the superior` accessibility of modular questions allied to somewhat lower boundary requirement for linear schemes, is consistent with that definition.

The debate about question structure is still in its infancy. Suitably structured questions may reduce the need for the employment of strategy in answering questions, and it is a matter of trust that awarders consider this. The practice of planning will enhance competence, as will practice in any particular type of question, and probably the value of learning strategic choice depends on its relevance to any further course or employment. It is probably true that there are more 'cues' in modular questions, already identified as affecting question demand, but equally, if similar types of question appear in linear schemes year after year, then practice of these questions will mean that strategy is a well learnt art for linear candidates for many of the questions. Unfamiliar questions, although well structured are likely to be as difficult to the candidate as familiar questions, albeit unstructured. There is, currently, a considerable amount of research being carried out in the field of question demand (see, for example, Pollitt et al (1998)) in various subject areas, although the focus is not on modular/linear schemes of assessment.

Despite the analytical research presented here, whilst there are some differences in the performance of questions within the two schemes of assessment, there is still no evidence to suggest that there is any major discrepancy in grading standards. There are inclines of difficulty in all the questions and some evidence of ceiling and floor effects, but not enough to undermine the effectiveness of the examination. The omission of negative findings might, therefore, suggest some comparability in the area of question difficulty, whilst recognising that there are some differences which could lead to discrepant performances. There is nothing which would imply that illegitimate variablilities had not been taken into account in the grading process, and some evidence that they have.

CHAPTER 8

If I Had but the Time - Longitudinal Analysis

In 1993 the first fruits of a DfE initiative, known as the 16+/18+ project, were seen. In essence this was the first matching of GCSE candidates to GCE candidates, and in the first year was based on a 10% sample. The rules of combination were very precise since only those who were 18 on 31st August 1993 were included in the matching process. However, this would normally include most of the candidates taking A levels in that year.

The original contract for this work extended until 1995 (and is still extant though under slightly different terms) and as a result (the 1996 dataset not being available yet) there are two full sets of matched data available for use in this study. These are the datasets for candidates taking A levels in 1994 and 1995, matched against their GCSE results, obtained mainly in 1992 and 1993 respectively.

It turns out that, for no obvious reason, most of the OCSEB linear mathematics (9650) candidates are missing from the 1994 matched data. Only 28 entries out of a total of 378 are held, which is too small a sample to be included in the initial phase of this longitudinal part of the study. However, the other OCSEB linear mathematics (SMP 9652) was well represented, and therefore will be used as the control for 1994, with both 9652 and 9650 being analysed alongside the modular scheme in 1995. It may be worth noting that two years later, in 1996, both the linear syllabuses were combined to form the one linear syllabus for OCEAC/OCSEB. Additionally, the SMP syllabus is tiered, with the syllabus 9655 examination a subset of that for 9652 and for which the maximum grade possible is C. An additional complication is that candidates were permitted to enter this examination through other Boards (a practice of sharing examinations which has largely discontinued) so the cohort is more varied that would be found for the traditional syllabus. For the purposes of this analysis only the 9652 results were used and it is important to stress that weaker candidates would have been entered for the limited grade syllabus, and would therefore not be represented. How analogous this is to modular candidates who choose not to complete the course or who withdraw from certification because of weak final modules is difficult to judge. The grade distribution for the all entries to 9652 in 1994 are given below (the comparable figures for 9665 and 9650 are given in chapter 5,

but are repeated here for reference). The total number taking the A level was 1684 (498 sat the restricted foundation examination).

Syllabus Distributions for 1994 - All Boards

	A	B	C	D	E	N	U	No
Cum % 9650	37.1	56.9	71.8	83.9	90.0	95.4	100.0	912
9652	34.0	55.6	71.7	84.4	90.4	95.4	100.0	1684
9665	33.3	58.2	78.0	91.3	99.0	100.0	100.0	4360

The other year of interest for this analysis is 1995, and below the grade distributions for the syllabuses 9665, 9650, 9652 are given.

Syllabus Distributions for 1995 - All Boards

	A	B	C	D	E	N	U	No
Cum % 9650	31.9	50.5	69.1	85.1	95.2	99.5	100.0	188
9652	44.5	63.6	77.9	89.1	95.0	98.3	100.0	1435
9665	28.3	53.3	73.6	89.0	98.2	100.0	100.0	6182

Although the total number certificated in 1995 is 1200 more than 1994, there has been a change in the distribution of those numbers, with the popularity of the modular scheme growing to the apparent detriment of the two linear versions (because centres are free to choose any Board it can only be surmised that a number may have chosen to take 9665 instead of the OCSEB linear schemes). Since, in line with our assumption, differences in grade distributions for ostensibly the same subject are due to differing abilities of the compared cohorts rather than differences in grading standards (especially within the same examining Board where many of the awarding personnel will be common to all syllabuses within the same subject), it appears that the 9652 candidates had demonstrated a higher level of achievement than in 1994, whereas the 9650 and modular candidates somewhat less. One conclusion might be that this could, in part, be the result of the demographic changes indicated by the entries, but equally there could have been some easing of linear grading standards between 1994 and 1995.

The Data

The relevant A level data were downloaded from the years 1994 and 1995 together with their matched mean GCSE scores from previous years. The 1994 data has 3477 candidates, which is nearly 80% of the full cohort. Of the 1994 9652 candidates, 63% are represented by the dataset. Whilst this is far from perfect, the fact that the exact provenance of the candidates is known (in terms of syllabus), is an improvement on some of the previous studies in this area which have aggregated syllabuses, sometimes indiscriminately. The relative figures are:

1994

		A	B	C	D	E	N	U
9652	Actual	573	363	272	213	102	84	77
	Dataset	380	223	164	143	65	88	
	%	(66)	(61)	(60)	(67)	(64)	(55)	
9665	Actual	1453	1089	867	577	340	44	1
	Dataset	1225	849	664	428	256	55	
	%	(84)	(78)	(77)	(74)	(75)		

In all cases the dataset at each grade is within 5% of the proportion expected given the total numbers. The rather odd figures relating to the N and U grades for 9665 are almost certainly the result of post hoc withdrawal from certification. Just over 80% of the linear syllabus are male, in the modular case the figure is 71%.

1995

		A	B	C	D	E	N	U
9650	Actual	60	35	35	30	19	8	1
		44	26	18	14	12	1	0
		(73)	(74)	(51)	(47)	(63)	(12)	

9652	Actual	638	274	206	160	85	47	25
	Dataset	471	210	139	116	66	33	22
	%	(74)	(77)	(67)	(73)	(78)	(70)	(88)
9665	Actual	1750	1546	1252	956	567	111	0
	Dataset	1535	1277	1005	782	430	109	0
	%	(88)	(83)	(80)	(82)	(76)	(98)	

For both 9652 and 9665 the dataset from the matched candidates is very representative of the whole cohort. For 9652, there is less than 1% difference in the percentages achieving each grade and for 9665 just over 1% difference. The picture is rather different for 9650. Because the entry is relatively small, a small number missing makes a large percentage difference in each grade, up to 10%, or in the representative ratio between the entry and the matched dataset. The results of the modelling must be interpreted with respect to this lack of representativeness.

Both the mean GCSE grade and syllabus grade have been transformed into numbers, which is always a debatable procedure with categorical variables, though commonly done. At GCSE level, 7 is A, 6 a B and so on down to 1 for a G. (A^* was not awarded until 1994). At A level, 10 equates to A, 8 to B, down to 2 for a minimum pass at E.

In 1994, for the traditional scheme, the mean GCSE score is 6.4 and the syllabus average grade 6.8. For the modular scheme, the comparative figures are 6.2 and 7.3. Simplistically this would imply that the modular A level was more leniently graded. (It should be noted that the conversion from grades to integers uses a different scale for GCSE than for A level. The difference in GCSE scores is approximately one fifth of a grade, at A level approximately one quarter of a grade).

In 1995, the equivalent figures were: for 9650 (the traditional scheme) the average mean GCSE score was 6.5 and the mean A level grade was 7.3 and these were repeated (to an accuracy of 10^{-2}) for the SMP syllabus. The modular scheme had figures of 6.1 for the average mean GCSE score and 6.7 for the equivalent A level figures. These data imply a somewhat closer set of results for linear and modular schemes than was the case in 1994.

The Model

The multi-level model used for analysing these data is the same for each syllabus, and utilises the available variables which, from simple regression analysis, would appear to be significant. The approach is somewhat different from that of chapter 6, which was an unconditional multi-variate model. Here the model is conditional and, not surprisingly given previous work in this field (Tymms and Fitz-Gibbon, 1991; Tymms and Vincent, 1994), the main explanatory variable is the mean GCSE score, which explains more of the variance in the final A level grade than anything else available. (A slight digression, but without this variable, the largest contributor to r^2 is actually the number of GCSEs obtained below grade C, although negatively correlated with A level grade i.e. the greater the number of below C GCSE grades, the lower the A level score). Other studies have also included the total number of GCSEs taken by each candidate (e.g. O'Donoghue et al, 1996) as part of the explanatory set of variables. However, it was clear on examination of the 1994 data, that it was incomplete in the total number of GCSEs taken. The matching of later data sets seems to have been more successful, but for the 1992/1994 data there were a number of missing GCSE data. Simple regression techniques also showed the number of GCSEs did not add to the power of the model. Although the data have been 'cleaned' to the extent that where there were no GCSE data available that candidate was removed, it was felt that a mean score was probably a fair representation of ability at this level even if calculated from only partial data. In the same circumstances the total number of GCSEs could be very misleading.

The following description relies heavily on Goldstein 1995, though the notation is slightly different. As explained in chapter 6, the final model developed towards the end of this chapter would require an unwieldy number of subscripts were not other notation introduced. Suppose we define a relationship between a candidate's mean GCSE score and their A level grade by the simple linear regression equation of A level grade on mean GCSE score as follows:

$$y_{ij} = \alpha_0 + \alpha_1 x_{ij} + e_{ij} \qquad (1)$$

where

y_{ij} = A level grade for candidate i in school j
α_0 = intercept of the regression line with the x-axis
α_1 = gradient of regression line
x_{ij} = mean GCSE score for candidate i in school j
e_{ij} = level 1 residual error

Whilst this model describes candidate (level 1) behaviour, it does not allow for group influences which might occur because of the school which the candidates attend. This centre level (level 2) variation is modelled firstly by allowing the intercept to vary and α_0 in equation 1 becomes $\alpha_0 + u_j$. This would be equivalent to different levels of performance within different centres, but keeping the relationship between the two variables constant, i.e. the set of regression lines would be parallel.

However, it is highly unlikely that the relationship between the input, or explanatory variable, and the response variable would be the same for all schools. Additionally therefore, the gradient of the regression line should also be allowed to vary between schools. Thus α_1 becomes $\alpha_1 + \beta_{1j}$ where j indexes the school.

$$y_{ij} = (\alpha_0 + u_j) + (\alpha_1 + \beta_{1j}) x_{ij} + e_{ij} \qquad (2)$$

where u_j and β_{1j} are random variables at centre level and, as before, e_{ij} is the residual value for the ith candidate at the jth school.

Rearranging into fixed and random parts we have:

$$y_{ij} = (\alpha_0 + \alpha_1 x_{ij}) + (\beta_{0ij} + \beta_{1j} x_{ij}) \qquad (3)$$

where $\beta_{0ij} = u_j + e_{ij}$

In order to take account of a possible non-constant variance of the level 1 residuals, it is necessary to allow for the coefficient of the GCSE mean score to vary at level 1. (An assumption of the model given by equation 1 is that the distribution of the level 1 residuals about each score is normal with constant variance. This would be violated with non-constant residual variance). A further level 1 term is added, $\gamma_{1ij} x_{ij.}$, where γ_{1ij} is random at level 1.

The equation now becomes

$$y_{ij} = (\alpha_0 + \alpha_1 x_{ij}) + (u_j + \beta_{1j} x_{ij} + \gamma_{1ij} x_{ij.} + e_{ij}) \qquad (4)$$

The variance components of this equation can now be defined such that the total level 1 variance is given by:

$$S^2_1 = \sigma^2_e + 2\sigma_{e\gamma} x_{ij.} + \sigma^2_\gamma x_{ij.}^2 \qquad (5)$$

where σ^2_e, σ^2_γ and $\sigma_{e\gamma}$ are the parameters associated with the random coefficients e_{ij} and γ_{1ij} i.e.

$$\text{var}(e_{ij}) = \sigma^2_e; \quad \text{var}(\gamma_{1ij}) = \sigma^2_\gamma; \quad \text{cov}(e_{ij}\gamma_{1ij}) = \sigma_{e\gamma};$$

(See Goldstein, 1995 for the statistical justification for this). Running the model (described later) showed that the level 1 variance was linearly dependent upon x_{ij} i.e. σ^2_γ is zero. The level 2 variance components will be:

σ^2_u the variance of the intercepts u_j across all j centres
σ^2_β the variance of the gradients β_{1j} across all j centres
$\sigma_{u\beta}$ the covariance of u and β_1

Two dichotomous variables are now added, one which defines the A level syllabus for which the grade has been awarded (i.e. either linear or modular) and one which allows the gender effect to be estimated. These are allowed to vary across centres as well as candidates. Therefore two additional terms are added, $(\alpha_2 + \beta_{2j} + \gamma_{2ij})s_{ij}$ where s_{ij} is the dummy variable for the syllabus type (0 for the linear syllabus, 1 for the modular form) and $(\alpha_3 + \beta_{3j} + \gamma_{3ij})\, g_{ij}$ is added where g_{ij} is a dummy variable (0 for a boy, 1 for a girl). Although it is entirely possible that gender has an effect at both candidate and centre level, it turns out that the only significant variance is at level 2, with all covariances non-significant. If we designate $u_j + e_{ij}$ as β_{0ij} for consistency then the full model split into fixed and random effects is given by:

$$y_{ij} = (\alpha_0 + \alpha_1 x_{ij} + \alpha_2 s_{ij} + \alpha_3 g_{ij}) + (\beta_{0ij} + \beta_{1j} x_{ij} + \gamma_{1ij} x_{ij.} + \beta_{2j} s_{ij} + \gamma_{2ij} s_{ij} + \beta_{3j} g_{ij} + \gamma_{3ij} g_{ij}) \quad (6)$$

With each random effect there will be an associated variance and co-variance, although some of these are not significant and will be removed.

Exploration of this model did suggest that not all possible combinations were contained in equation 6, and it was apparent that further fixed effects were sometimes significant. Terms involving higher powers of the mean GCSE score, i.e. x^2_{ij} and x^3_{ij} become additional as $\alpha_4 x^2_{ij}$ and $\alpha_5 x^3_{ij}$ in the fixed part of the equation above and the compound term of GCSE score with syllabus with coefficient α_6 was also significant. The final model equation now becomes:

$$y_{ij} = (\alpha_0 + \alpha_1 x_{ij} + \alpha_2 s_{ij} + \alpha_3 g_{ij} + \alpha_4 x^2_{ij} + \alpha_5 x^3_{ij} + \alpha_6 x_{ij} s_{ij}) + (\beta_{0ij} + \beta_{1j} x_{ij} + \gamma_{1ij} x_{ij.} + \beta_{2j} s_{ij} + \gamma_{2ij} s_{ij} + \beta_{3j} g_{ij} + \gamma_{3ij} g_{ij}) \quad (7)$$

The Results - 1994

The data for the two syllabuses 9652 (linear) and 9665 (modular) were combined for analysis. The first step in the process was to look at the results from standard, single

level regression analysis on the raw, untransformed data. This resulted in the following equation:

$$y_{ij} = 17.92 - 10.16\, x_{ij} + 1.8\, x^2_{ij} - 0.07\, x^3_{ij} + 4.09\, s_{ij} - 0.91\, g_{ij} - 0.45\, x_{ij}\, s_{ij}$$

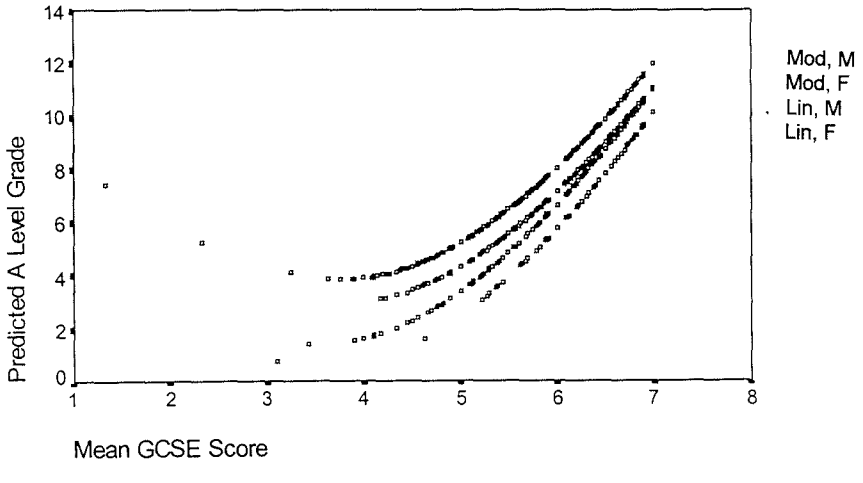
where for candidate i in centre j.

y_{ij} = A level grade
x_{ij} = mean GCSE score
g_{ij} = gender dummy, 0 for boys, 1 for girls
s_{ij} = syllabus dummy, 0 for linear, 1 for modular

The residual variance is 4.5 and the model explains some 44.0% of the total variance. In fact there are a number of different combinations of variables which produce similar results, but none of them explain more than 44% of the variation.

The plot of the predicted values of y_{ij} , the A level grade against the mean grade (x_{ij}) is given in figure 8.1. The four distinct curves equate to the four pairs of settings of the dummy variables (1,0; 1,1; 0,0; 0,1) for the syllabus and gender data. The curvilinear effect is produced from the addition of the non-linear term and towards the lower end of the mean GCSE range the differences in the predicted A level grade between the four settings increases. At the top end of this range, the results for linear boys is almost coincidental with those for modular girls, but the converging pattern of all four curves is discernible.

Figure 8.1: Predicted A level Grade from Simple Regression



Also, below GCSE scores of 3 for modular boys, the regression equation would appear to predict an increase in A level grade with decreasing GCSE score. This is undoubtedly due to the sparse and untypical nature of A level scores at the bottom of

the mean GCSE range. These are unacceptable results and throw into relief the problems attendant upon too simplistic an approach.

None of this is particularly surprising, but further analysis using multi-level modelling, which takes into account variation at centre level, indicates that there are other effects that are not apparent from the single-level analysis. Although the majority of the rest of this chapter uses transformed data, the inclusion of the following, initial, untransformed analysis is an aid to understanding. Transformations can sometimes produce enhanced and undesirable effects and it is important to gauge output from the model initially using the more familiar untransformed raw data.

Extending the regression model to account for a second level of variation, (where level 1 variation is between individual candidates and level 2 between centres), there are four explanatory variables, a constant vector associated with the intercept α_0 in equation 6, the mean GCSE score (x_{ij}) associated with the fixed vector α_1 and the binary variable s_{ij} representing the A level mathematics syllabus taken by the candidate and the dummy variable g_{ij} representing the gender dummy variable, set to 0 if male and 1 if female. There is also the interaction of mean GCSE score with syllabus to be included, whose coefficient is denoted by α_6.

The results from the initial analysis are given by:

Parameter	**Estimate**	**Standard Error**
Fixed		
α_0 (cons)	-0.55	2.62
α_1 (meang)	-1.12	0.79
α_2 (syll)	4.82	1.23
α_3 (sex)	-0.77	0.09
α_4 (meang^2)	0.36	0.06
α_6 (meang*syll)	-0.59	0.18

with random elements given in the following variance/co-variance matrices:

Random

Level 1

	e	γ_1	γ_2	γ_3
e (cons)	44.8 (4.16)	-3.03 (0.31)	-8.58 (2.12)	-5.25 (1.13)

γ_1 (meang)	0	1.08 (0.31)	0.86 (0.17)
γ_2 (syll)		0	-0.11(0.31)
γ_3 (sex)			0

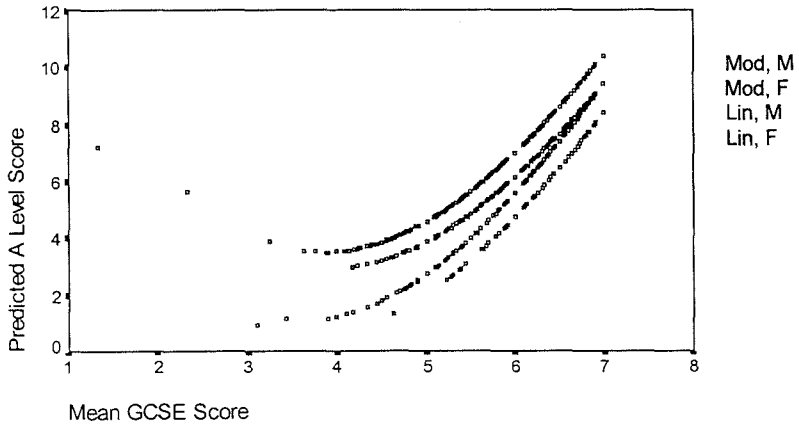
Level 2

	u	β_1	β_3
u (cons)	9.04 (2.92)	-1.26 (0.43)	1.12(0.5)
β_1 (meang)		0.19 (0.07)	-0.21(0.08)
β_3 (sex)			0.54 (0.16)

There are a number of interesting points to notice, but of most relevance to the current exercise is that none of the variance at centre level is explained by the syllabus taken. This is slightly surprising since it would be expected that the influence of centres on the results from modular schemes could be greater because of the interactive nature of the scheme even if only considered in the timing of entries and resit policy. One would expect on a greater variance between centres as that influence (not applicable in linear schemes) is exercised, or not. However, such an effect is not seen when the two syllabuses are investigated as part of one dataset, as here.

The fixed effects are plotted on figure 8.2.

Figure 8.2: Predicted A level Grade from Fixed Effects



Comparison with the results from the initial regression analysis (figure 8.1) show that the results are very similar, especially at the top end of the mean grade range. There is somewhat more spread, however, at the lower end. There are two major points of

interest; one of which has been used by other studies to confirm suspicions that modular schemes are easy, and one that has apparently been undetected.

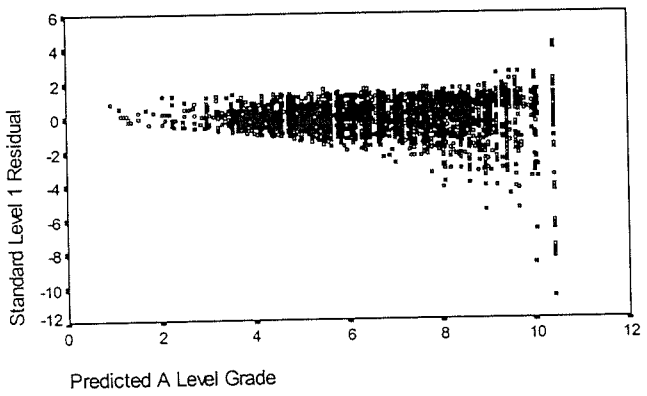
(a)for the same mean GCSE score male linear candidates would appear to be obtaining lower grades than their modular counterparts. Similarly female linear candidates are apparently gaining less reward than their GCSE scores might suggest compared with their modular counterparts. However, at the top end of the ability range male linear candidates appear to be gaining equal reward to that of female modular candidates.

(b)the curves are converging towards the top end of the mean grade range implying that any difference decreases with GCSE attainment.

There is, therefore, evidence that modular candidates are performing somewhat better that their GCSE scores might predict. This suggests that either grading standards in the two schemes are different thus violating the assumption of comparability or that modular candidates are 'over-performing' (or possibly linear candidates are 'under-performing'), especially at the lower end of the ability range, or a combination of the two.

The level 1 residual plot (figure 8.3) is symmetrical, but not constant. This suggests that the application of a suitable transform might reasonably be used in order to stabilise the variance of residuals.

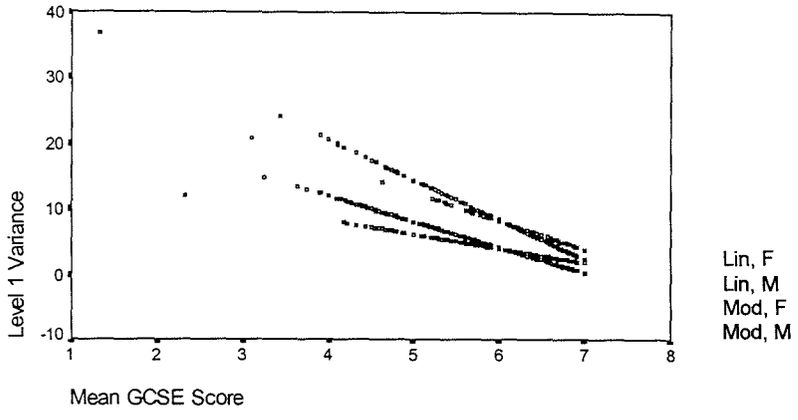
Figure 8.3: Standardised Level 1 Residual Plot



The level 1 variance plot (figure 8.4) shows the four distinct lines of the dummy pairs, but the interesting feature is the cross-over of the pairs at a mean score of about 6. At low scores, the top two lines are the variances of the linear syllabus for boys (top)

and girls, the bottom two are for the modular scheme boys and girls (lowest). At the maximum average score, this order has changed such that the highest spread in residual variance is for linear females, while the lowest is that for males taking the modular examination, with the linear boys and modular girls showing very similar intermediate variances.

Figure 8.4: Level 1 Variance



The lowest level 1 variance of 0.45 at the top end of the predicted grade range is for modular males. Since the maximum predicted A level grade is just under 10, it is clear that the variance will not be symmetric about this prediction because a normal residual distribution with a standard deviation of 0.67 would contain data points which exceed this value. The largest residual variance in maximum predicted score is 3.81 for linear females, which again will not be symmetric since despite the predicted maximum score being somewhat lower at just over 8, a standard deviation of 1.95 would again imply values above 10 should the distribution be normal about the prediction. A transformation would go some way to alleviating this problem.

For candidates obtaining an average GCSE grade of C, the spread of predictions for their A level grade is somewhat greater. It would be expected that 95% of the linear boys would obtain a result within 4 grades of the predicted grade, whereas the comparable range for modular girls would be just under 2½.

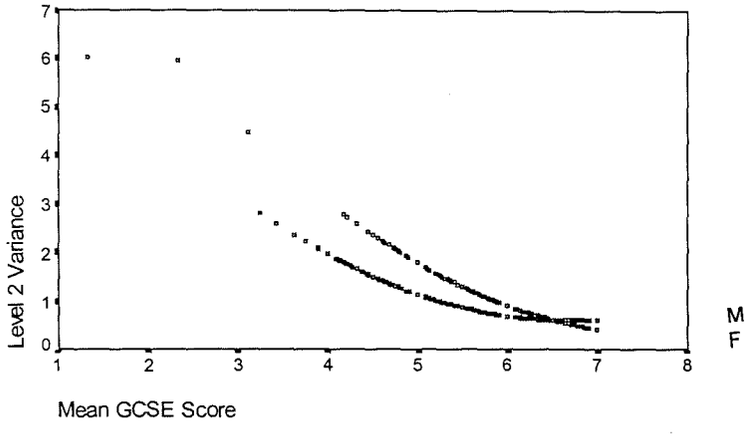
There are a number of observations which follow from these results. Firstly, and most obviously, is that the level 1 variance is not constant. It is dependant on both syllabus and gender and is a function of the mean GCSE score. The effects of the interaction of these factors not negligible.

For example, GCSE mean scores are better predictors of attainment in A level mathematics examinations when these are assessed in a modular mode than for those which are linearly assessed because of the smaller variance associated with this syllabus. There may be an implication that modular examinations are more GCSE-like, or possibly that they allow candidates better to fulfil potential demonstrated at GCSE, but there is an outcome which is less tenuous. The greater spread of the best linear candidates would mean that some high GCSE scoring candidates would get lower A level grade and the average of this would produce the fixed effect shown in figure 8.2, including the convergence. There has also been mention in the past that there is a "ceiling effect" which attaches to modular schemes. (This has been alluded to in the previous chapter on question analysis.) There is a possibility that because of its nature and question structure, the modular examination, whilst designed in part to enable weaker candidates to benefit from the course, leaves some candidates "under-extended". Although the ease, or otherwise, with which the best candidates perform is untestable (at least within the terms of this thesis), it is certainly true that they cannot get better than an A. However, the converse is that weaker candidates may get little out of studying for the linear scheme since the range of predicted A level grades (which the fixed effects show to be about one grade lower) at this end of the ability range is greater. But it is the dilemma in debates on the relative merits of assessment methods that what, on one side, is leniency, on the other is proof of the merits of the assessment method in question.

Also there is the gender effect which indicates that the predictability of A level grades is better for girls at the lower end of the ability range, and for boys at the higher end. A possible explanation is that the generally better performance of girls (Arnot et al, 1996) gives their mean GCSE grade a predictive capability which may not attach to low scoring 'lazy' boys with ability. However, for high attaining boys who do well at GCSE, their results are mirrored by A level performance. There is also the suspicion that high GCSE performers amongst the girls are more likely to underperform at A level than their male counterparts, so giving rise to the greater spread of results. There is certainly some evidence which supports this view (Equal Opportunities Report, 1995). The gender-syllabus interaction was not a significant effect.

The strength of this type of analysis is its ability to look beyond the simple and consider the variation between centres. The level 2 variation is illustrated in figure 8.5. It indicates that there is more variance in girls' results at centre level, but that, as previously mentioned, this is unaffected by the syllabus taken by the candidates. The dependency of the level 2 variance on GCSE scores results in a considerable increase in this variance as GCSE scores decrease. In other words the centre level effect is magnified for weaker candidates. This is probably a fairly intuitive result: good, probably well motivated, candidates do well anyway, it is with the weaker brethren that centres can have more profound influence. Certainly it would appear so for these schemes. However, there is also a possible scale effect and a transformation may be desirable in order to mitigate this.

Figure 8.5: Level 2 Variance



There is also a problem with the estimate of σ^2_u, the variance of the intercept across centres. As the ordinate is currently specified this variance would be when the mean GCSE score was zero - not a likely situation. The usual method of "centring'" the variances is to move the ordinate by a simple linear transformation to the average score. This centres the regression lines rather nicely as well as improving numerical stability, and the results from the model are still easy to interpret in terms of the initial variables. But there are reasons why other transformations may be preferable, desirable statistical properties can be built into the transformed data in order to aid analysis and provide rigour to the modelling.

It has been suggested that a more useful transformation for this type of analysis is to normalise scores of the mean GCSE grade and especially the distribution of A level grade, again centred on a mean of 0 and a standard deviation of 1 (O'Donoghue,

Thomas, Goldstein and Knight, 1997). One effect is to reduce the influence of certain parameters while enhancing others and the model is often simplified.

The choice of transformation is not always obvious. The outcome of a systematic alteration in the variables must have desirable characteristics which aid either analysis or interpretation of results to be of any value. Centring scores so that they have a mean of 0 but are otherwise unchanged might help in scaling (as in the case of the mean GCSE scores) and stability but does little to aid analysis as all other distributional properties are unchanged, most specifically the variability and standard deviation.

Two other types of transformation are commonly used in this situation. The first, a transformation to standard scores, results in a distribution which has a mean of 0 and a standard deviation of 1. However, the shape of the frequency distribution is the same i.e. the skewness, kurtosis and rank order are unchanged (this latter is an essential requirement of any transformation in the current analysis) as is the proportionality of scale.

Because the grading process is based on judgements and categorises performances, the relationship of grades to each other (in terms of raw marks) may seem fairly arbitrary and is an artefact of the mark scheme and aggregation process of the examination. There is thus little of fundamental importance, except the rank order, which needs to be preserved in order to investigate the statistical relationships of the type considered here. The most obvious transformation to use is that of Normalisation of the scores. This distributes the grades normally, in this case with a mean of 0 and a standard deviation of unity.

Although easy understanding of the results is affected by normalisation, the benefits almost always outweigh this drawback. Many statistical tests assume normality, and are more robust if such can be demonstrated to be the case. Additionally the normality of the compared distributions (in this case A level score and mean GCSE score) may reduce the number of parameters needed to describe their variability. This is undoubtedly an asset in multi-level modelling when fewer variables may need to be included to examine the behaviour of the data.

Another useful technique in defining the model is the use of a spline function. Whilst the power of the model may be enhanced by the use of normalised distributions, a side effect is that the contribution to the regression equation of the tails of the distribution are exaggerated with the result that this unrepresentative data leads to rather unlikely looking outcomes - in this case the addition of the quadratic and cubic terms leads to an apparent upturn in the fortunes of candidates with very low GCSE scores. Because it is an unrealistic output from the model, it is one which any transformation should aim to overcome. Since both the simple regression model and the multi-level model using untransformed data indicate a similar, but unacceptable, pattern, there needs to be a smoothing operation to ensure that the model accurately reflects the behaviour of the data in those ranges where the it is truly representative, but reduces the influence of clearly untypical data points. The solution is to replace both terms with a spline function such that the spline $(x - t)_+^n$ has the property

$$(x - t)_+^n = (x - t)^n \quad \text{for } x > t, \quad n = 2, 3 \text{ etc.}$$
$$= 0 \quad \text{for } x \leq t$$

In this case a suitable value for t is -2 i.e. the value for x at which the model becomes unreliable. This essentially constrains the higher power terms in the regression equation to that part of the data which is most representative, so that the fixed part of the model in (7) becomes:

$$\alpha_0 + \alpha_1 x_{ij} + \alpha_2 s_{ij} + \alpha_3 g_{ij} + \alpha_4 (x_{ij} - t)_+^2 + \alpha_5 (x_{ij} - t)_+^3 + \alpha_6 x_{ij}\, s_{ij} \qquad \text{.....(8)}$$
$$\text{where } t = -2$$

Variances are largely unaffected by the use of the spline, though cubic and quadratic fixed effects are dampened at the tail.

Using the normalisation transform for the response, A level grade, and the mean GCSE score (meangn), the significant model estimates are:

Parameter	Estimate	Standard Error
Fixed		
α_0 (cons)	-0.74	0.09
α_1 (meangn)	0.38	0.07
α_2 (syll)	0.34	0.05
α_3 (sex)	-0.26	0.03
α_4 (spline^2)	0.13	0.02
α_5 (spline^3)	-0.0005	0.00005
α_6 (meangn*syll)	-0.09	0.03

Random

Level 1

		e	γ_1	γ_2	γ_3
e	(cons)	0.62 (0.03)	-0.13 (0.01)	-0.15 (0.02)	0
γ_1	(meangn)		0	0.04 (0.01)	0.06 (0.01)
γ_2	(syll)			0	0
γ_3	(sex)				0

Level 2

		u	β_1	β_3
u	(cons)	0.06 (0.01)	0	-0.02 (0.01)
β_1	(meangn)		0.01 (0.001)	-0.01 (0.001)
β_3	(sex)			0.04 (0.01)

The total level 1 variance will be given by:

$$S_1^2 = \sigma^2_e + 2\sigma_{e\gamma_1} x_{ij.} + 2\sigma_{e\gamma_2} s_{ij.} + 2\sigma_{\gamma_1\gamma_2} x_{ij} s_{ij} + 2\sigma_{\gamma_1\gamma_3} x_{ij} g_{ij.}$$

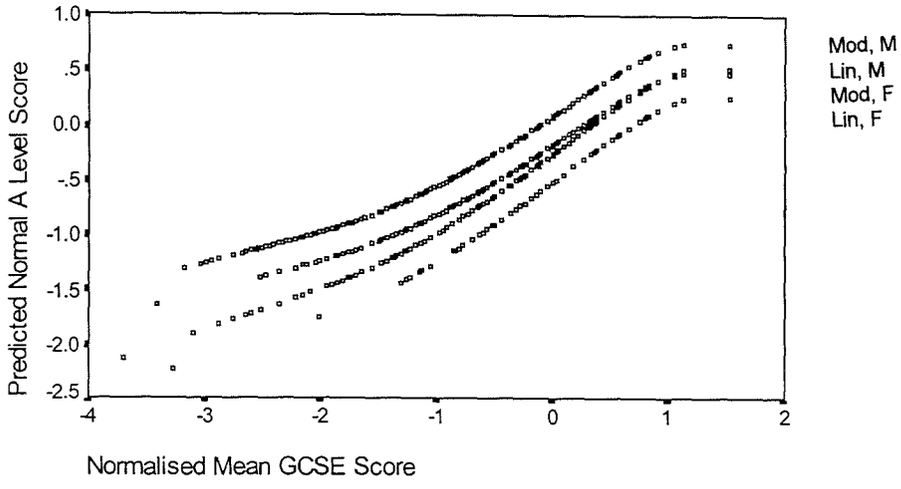
and the level 2 variance:

$$S_2^2 = \sigma^2_u + 2\sigma_{u\beta_3} s_{ij} + \sigma^2 \beta_1 x_{ij}^2. + 2\sigma_{\beta_1\beta_3} x_{ij} g_{ij} + \sigma^2 \beta_3 g_{ij.}^2$$

In addition to the fact that these values are an order of magnitude smaller than those from the first analysis, there is little change in those factors which are significant in the fixed effects part of the model, with the spline allowing the inclusion of both quadratic and cubic terms, although the influence of the latter is reduced.

The estimate of the fixed effects suggest that those related to the syllabus, candidate gender, mean GCSE normalised score and the spline quadratic and cubic functions of the normalised mean GCSE score are all highly significant, although to somewhat different degrees. The cubic effect is smaller than that from the quadratic term, but is instrumental in defining (smoothing) the curve at the top end of the ability range. This is consistent with intuition where expectation would be that candidates with less than perfect, though still very high, average GCSE scores would attain an A at A level. The maximum coefficient in this conditional model is still the explanatory variable associated with the mean GCSE score, although here normalised. The fixed effects estimates have been plotted in figure 8.6.

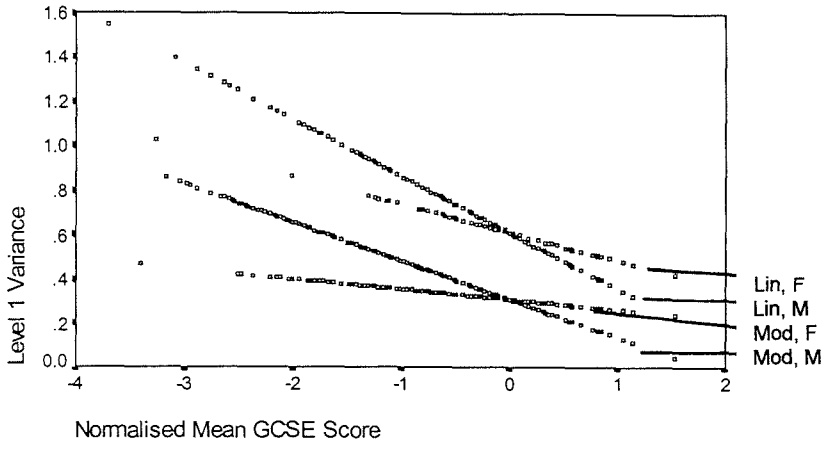
Figure 8.6: Predicted Normalised A Level Grade from Fixed Effects



The non-linear form of the curve is somewhat more marked than in the previous analyses and the ceiling effects more apparent. However there is still essentially the same pattern as previously found (figure 8.2) which suggests that for a given mean GCSE score, although a candidate may be likely to get a higher grade from the modular examination than from the linear scheme of assessment, the gender effect is nearly as great. At the top end of the ability range, linear males just outperform females taking the modular scheme. Again, the two possible explanations lie in inequality of grading standards or in the facility of the modular curriculum and assessment to improve the level of attainment of candidates.

The pattern of variances is again similar to those from the untransformed model. At the candidate level (level 1) there are again two clear pairs of bisecting lines, shown in figure 8.7. These correspond to the two syllabuses (the top pair illustrating the linear syllabus) and show that for both syllabuses the variance in A level grades is greater for females at the higher GCSE grades, for males at the lower end of the scoring range. At the average value for the mean GCSE score, under the normalisation transform, the variance for the linear scheme is about twice that for the modular assessment. This indicates that as a predictor for the A level grade, the mean GCSE score is more effective for the modular scheme.

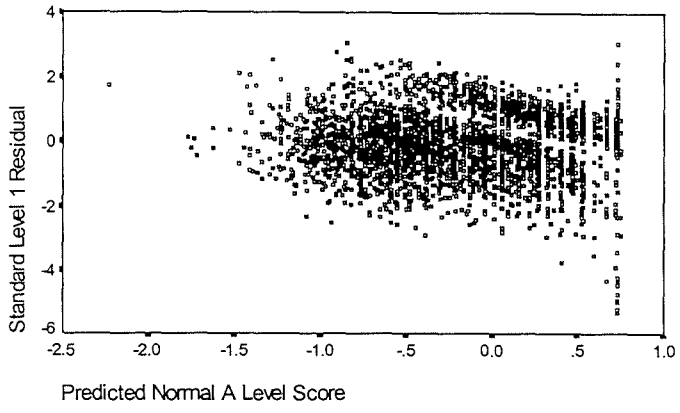
Figure 8.7: Level 1 Variance



It may be that the assessment method of modular schemes (or at least this particular mathematics scheme) make them more 'GCSE like' than the more traditional format. The somewhat gentler approach to the actual written examination with its associated formative and diagnostic assessment, shorter and more focused examinations may add some credence to this idea. The cognitive aspects of such connections are, however, at best, rather tenuous, and perhaps too complex for mere conjecture. What may be concluded is that less reliance can be put on the predicted value of the A level grade for the linear scheme given the mean GCSE score.

The plot of the level 1 residuals, (which have been standardised in figure 8.8) apart from scale, shows that the transformation has resulted in a more even spread of residuals throughout the range of predicted scores. Despite normalisation there is still evidence of a ceiling effect at the top end of the ability range, an effect which is probably enhanced by the normalisation of the relatively crude of the A level grade scale.

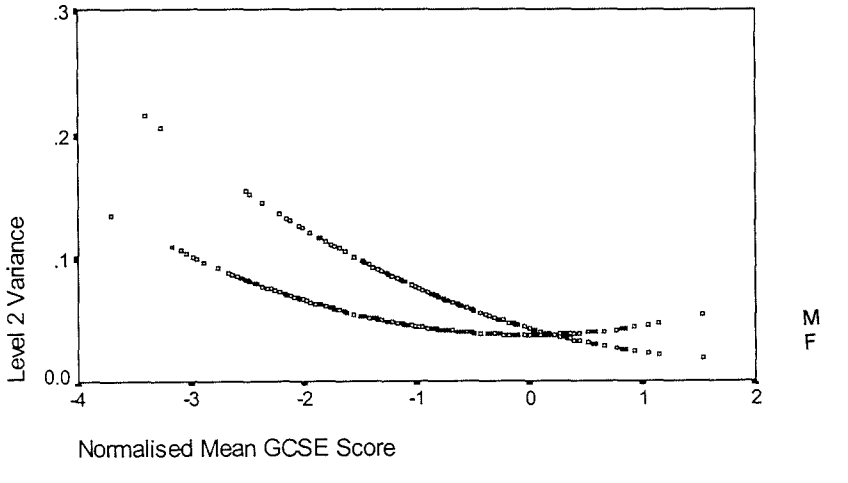
Figure 8.8: Standardised Level 1 Residuals



Results from the model define an envelope within which 95% of the predicted scores might be expected to lie given the calculated level 1 variance (i.e. ±1.96 SD). All the predicted values are well within these limits, irrespective of gender or syllabus. This suggests that at level 1 the null hypothesis holds i.e. there is no significant difference between predicted A level scores of the defined groups of candidates.

Of equal interest is the picture given of the level 2 variances (figure 8.9). There is again a clear intersection in the variances between boys and girls, and here also there is no significant syllabus effect.

Figure 8.9: Level 2 Variance

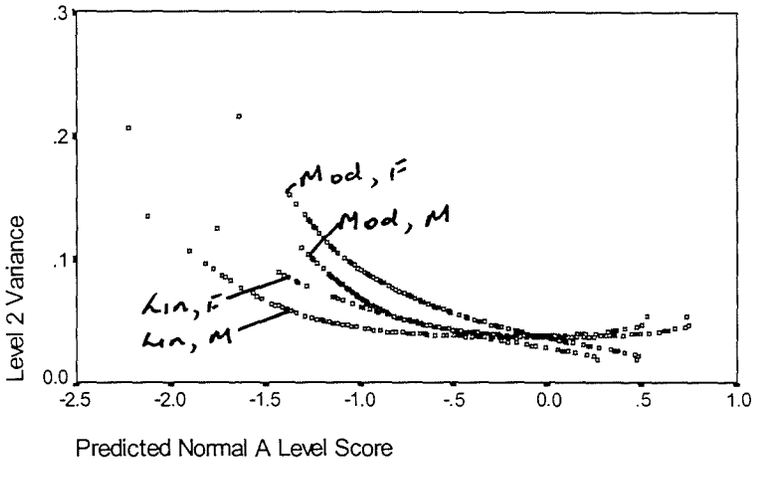


Additionally, at the centre level, these variances are non-constant. At low mean GCSE normalised scores, the level 2 variance for females is greater than that for males and repeats the picture given in the first analysis. However, above normalised scores of about 0 (which would be the average mean GCSE score) there is a complete reverse and the variance in predicted normalised A level scores at centre

level is greater for boys taking the modular scheme. This is because under the transformation the covariance of the gender variable with normalised score is a significant factor in the calculation of the total variance at centre level. This an effect not seen in the untransformed analysis and is almost certainly due to the definition given to the top candidates under the transformation. The implication is that although girls, especially the less able, are generally more likely to be influenced by the centre and teaching they receive, at the very top end of the ability range it is boys who are affected most, and the least affected group overall are the middle range candidates.

Intuitively one would expect there to be some difference between level 2 variances for the two schemes not only because there is a greater variety of centre types participating in the modular scheme but also because the assessment scheme allows centres to have more input in deciding a number of issues. A slightly different picture emerges when the level 2 variance is plotted against the predicted A level score (figure 8.9a).

Figure 8.9a: Level 2 Variance as a Function of Predicted A Level Grade



Because of the syllabus input to the fixed effects part of the model, the variance of those predictions between centres will depend upon which syllabus is taken. What emerges is a picture which suggests that the variance in predicted A level scores between centres is greater for modular schemes than for linear schemes, with again a greater variance for females than males at the lower end of the ability range. This result endorses the view that between centres it is easier to predict accurately results from the linear scheme. Almost certainly the influence of centre type (which is far more diverse for the modular syllabus) is one reason, but the effect of the scheme itself must be another.

The variance analysis suggests that fixed effects should be treated with caution, especially when combined with the syllabus and gender effect which, as shown in figure 8.2, confuses any clear distinction between linear and modular schemes and certainly must cast doubt on any unequivocal conclusions. The variances at both candidate and centre level mean that any between syllabus differences in predicted scores are not significant.

The Results - 1995

A similar analysis was carried out on the 95/93 matched dataset, but the issues for investigation were slightly different. Firstly it was necessary to attempt to establish whether results for the linear 9650 candidates could be extracted and extrapolated to the 1994 data. Unfortunately it proved impossible to separate out this syllabus explicitly in the multi-level analysis. Not only did all differences between this traditional syllabus and the SMP syllabus prove non-significant, but the small numbers proved insufficient to produce any other significant effects. Therefore both the linear syllabuses have been combined for the purposes of analysing the 1995 dataset.

It was also important to establish whether the results from 1994 were reproducible with the new dataset. The factors which proved significant in the analysis of the 1994 data and the results from the regression analysis need not necessarily have proved to be the same year-on-year. In fact there was broad agreement in those factors which explained the variance, and also in the pattern of both the fixed effects and random parameters.

For the syllabuses investigated here there seems to have been more success in the matching process as the percentages within each grouping indicate. This may just be a result of experience of matching rather than anything more profound. Only three factors proved of significance in the simple linear regression equation. These were the mean GCSE score, the gender dummy and the modular dummy. The resulting equation, which accounted for 44% of the variance is:

$$y_{ij} = 9.49 - 6.25\, x_{ij} + 1.24\, x^2_{ij} - 0.05\, x^3_{ij} + 0.58\, s_{ij} - 0.73\, g_{ij}$$

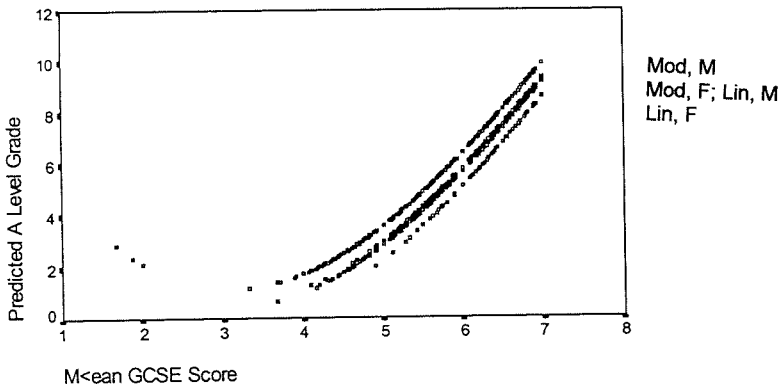
where

y_{ij} = predicted A level grade
x_{ij} = mean GCSE score

g_{ij} = gender dummy, 0 for boys, 1 for girls
s_{ij} = syllabus dummy, 0 for linear, 1 for modular

There are fewer terms than the equation produced by the 1994 data, and none that are compound implying no significant interaction between the mean GCSE score and either gender or syllabus type. The amount of variance accounted for in each case is similar. Comparison of figure 8.10 with the equivalent figure (8.1) obtained from the 1994 data shows a much more parallel picture for the four pairings of syllabus and gender, with an almost indistinguishable difference in predicted grades for the modular females and linear males (again as before).

Figure 8.10: Results from Simple Regression Analysis

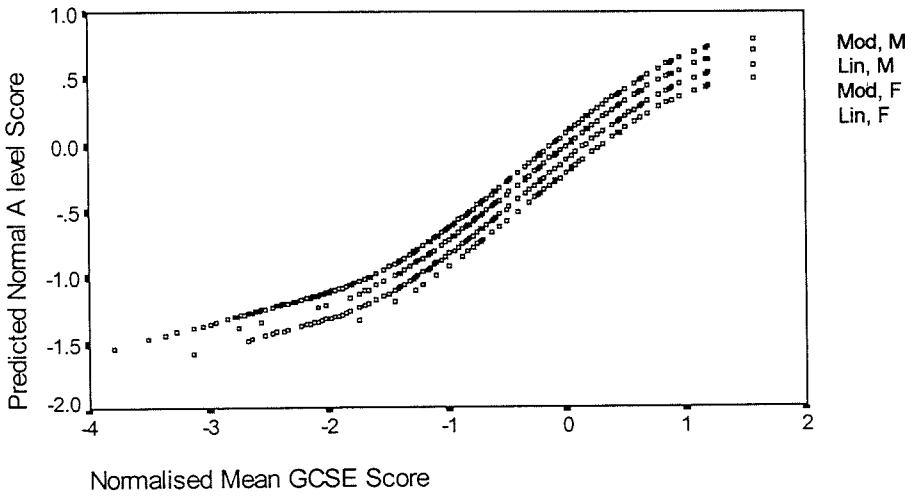


The multi-level analysis was carried out on the normalised mean GCSE scores and A level grades. Again a spline function was employed to model the non-linear fixed effects because of the exaggerated effect of these terms on low scores under the normalisation transform. The fixed effects were then estimated to be:

Parameter	Estimate	Standard Error
Fixed		
α_0(cons)	-0.73	0.11
α_1 (meangn)	0.24	0.06
α_2 (syll)	0.09	0.04
α_3 (sex)	-0.20	0.02
α_4 (spline^2)	0.30	0.04
α_5 (spline^3)	-0.06	0.01

Again comparison with 1994 shows that there are no interaction effects between the GCSE score and syllabus. This reinforces the results from the simple regression which indicated a far more consistent picture of predicted grades for each syllabus.

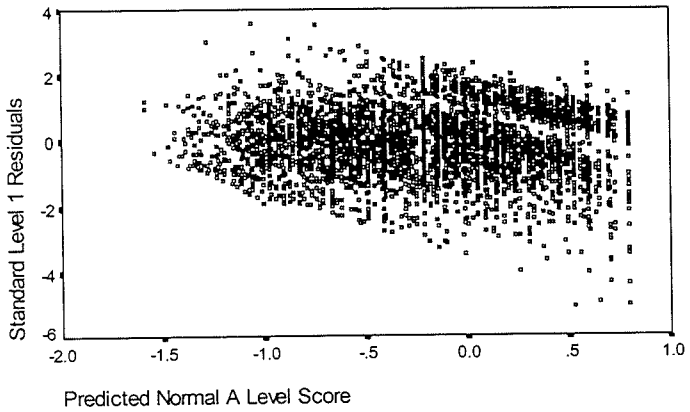
Figure 8.11: Predicted Normalised A Level Grade from Fixed Effects



The syllabus effect is much reduced as shown in Figure 8.11 which in general pattern bears comparison with figure 8.6, but again there is a greater degree of parallelism in the 1995 data. However, there are again the four lines representing the 0,1 dummies for syllabus and gender, with the upper pair showing the relationship between the GCSE score and the predicted A level for modular and linear males. The fact that the gender effect has not changed by much from 1994 indicates just how much closer together the two types of assessment scheme have become.

The standardised level 1 residual picture is shown in figure 8.12. It is very like that for 1994 suggesting equivalent accuracies of prediction throughout the predicted A level range, with again the suggestion of a ceiling effect.

Figure 8.12: Standardised Level 1 Residuals



The random part of the model gave the following estimates at level 1 and 2:

Random

Level 1

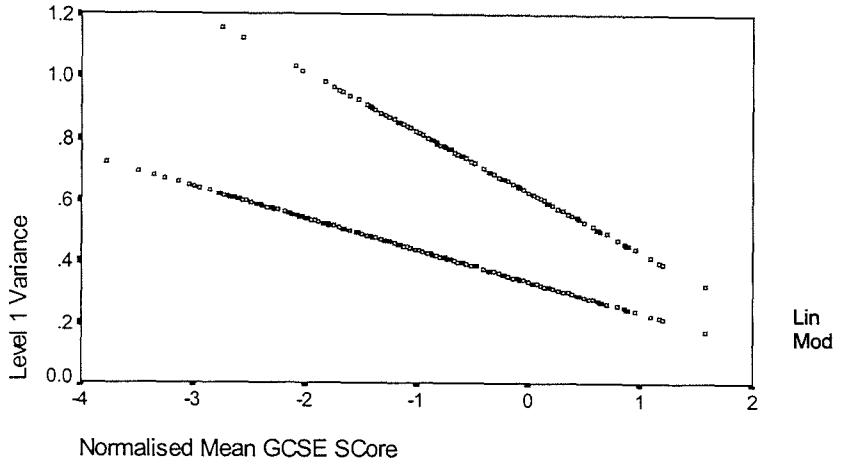
	e	γ_1	γ_2
e (cons)	0.63 (0.03)	-0.10 (0.01)	-0.15 (0.02)
γ_1 (meangn)		0	0.05 (0.01)
γ_2 (syll)			0

Level 2

	u	β_1	β_2	β_3
u(cons)	0.05 (0.01)	-0.01 (0.00)	0	0
β_1(meangn)		0.01 (0.00)	0.02 (0.01)	0
β_2(syll)				0
β_3(sex)				0.02 (0.01)

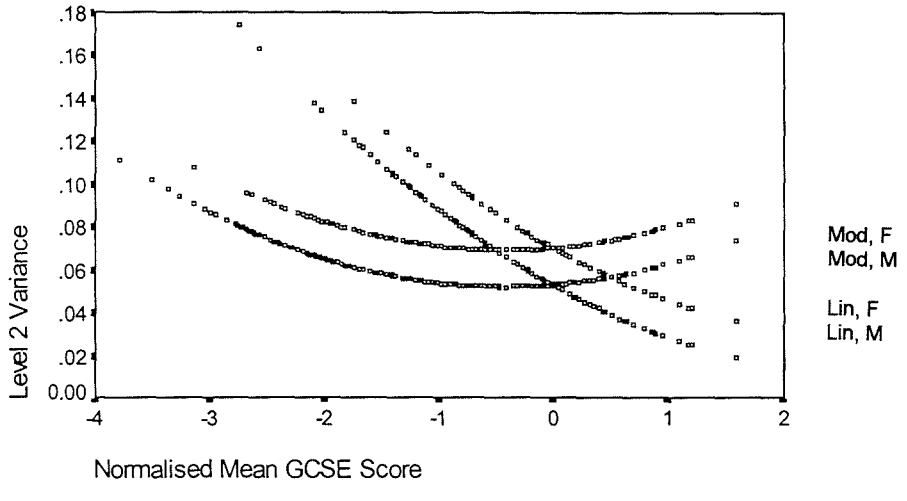
The level 1 variance (plotted in figure 8.13) shows only two lines (from the syllabus variable) since there would appear to be no gender dependence at this level.

Figure 8.13: Level 1 Variance



The lines approximate to the means of each pair of bisecting lines in figure 8.7. Again there is an implication that whatever differences were apparent in 1994 were very much reduced a year later. The confidence intervals defined from the level 1 variance reinforce this picture with all four prediction lines within the 95% limits suggesting that any differences are below significance at level 1.

Figure 8.14: Level 2 Variance



The picture given from the level 2 variances (figure 8.14) is different from that of 1994. There is a distinct syllabus effect (and this is evident whatever the abscissa chosen - either mean normalised GCSE score or predicted A level score) with the variance between centres lower for high ability children (and higher for females than for males) for the traditional scheme but at much lower than average GCSE scores, the between centre variance is higher for the traditional schemes.

This syllabus effect may be partially explained by the demographic changes between the years when much of the tail was lost by the traditional syllabuses and loss of data points may have led to a degradation in the regression coefficients. However, it may also represent a considerable learning curve on the part of centres who were all more able than in 1994 to use the flexibility of the modular schemes to their advantage. It would certainly seem that the change in the level 2 variance occurred primarily to the linear scheme, because that of the modular scheme seems relatively stable between the years.

Combined Data

The final analysis gives a composite picture from the combined data sets from 1994 and 1995. Level 1 is still candidate level, but level 2 becomes year and level 3 centre. The model for this analysis is very similar to that of equation 7 with an additional level fitted such that the subscript i denotes candidate, j year and k centres. The fixed effects part of the model will be unchanged, but the random part will contain additional terms needed to model the variance at level 2. Using the notation that $u_k + v_{jk} + e_{ijk} = \beta_{0ijk}$, we have for candidate i in year j and centre k becomes:

$$y_{ijk} = [\alpha_0 + \alpha_1 x_{ijk} + \alpha_2 s_{ijk} + \alpha_3 g_{ijk} + \alpha_4 (x_{ijk} - t)^2_+ + \alpha_5 (x_{ijk} - t)^3_+ + \alpha_6 x_{ijk} s_{ijk} + \alpha_7 z_{ijk}] + [\beta_{0k} + \beta_{1k} x_{ijk} + \delta_{1jk} x_{ijk} + \gamma_{1ijk} x_{ijk.} + \beta_{2k} s_{ijk} + \delta_{2ik} s_{ijk} + \gamma_{2ijk} s_{ijk} + \beta_{3k} g_{ijk} + \delta_{3jk} g_{ijk} + \gamma_{3ijk} g_{ijk}] \quad \text{.....(9)}$$

where δ_{1jk}, δ_{2jk} , δ_{3jk} and v_{jk} are the level 2 coefficients for mean GCSE score, syllabus, gender and constant respectively and + defines the spline functions. The most effective value for t is again -2. There is also an extra term in the fixed part of the model, $\alpha_7 z_{ijk,}$ which represents the term explaining the behaviour between years such that z_{ijk} is 0 when j is 1994 and 1 when j is 1995.

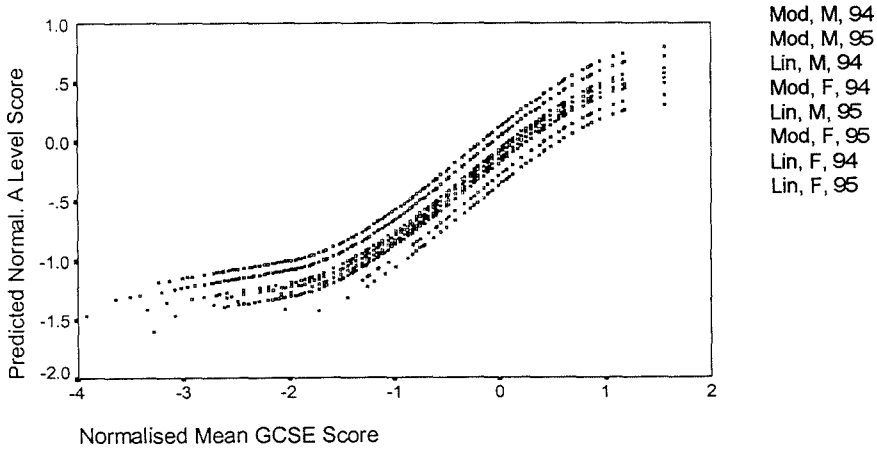
Running the model on the transformed mean GCSE scores and A level grades, normalised across the two years of data, produces the following parameters from the fixed part of the model:

Parameter	Estimate	Standard Error
Fixed		
α_0 (cons)	-0.90	0.09
α_1 (meangn)	0.14	0.05
α_2 (syll)	0.19	0.04
α_3 (sex)	-0.22	0.02
α_4 (spline^2)	0.35	0.03
α_5 (spline^3)	-0.07	0.00
α_7 (year)	-0.08	0.02

Apart from the value of the constant (intercept) term, these parameters are, in general, closer to those found for the 1995 data alone with a gender effect slightly greater than that for the syllabus. However, the offset explaining the difference between the years is fairly small and suggests a slightly worse performance in 1995, a result which might have been expected given previous analysis. Whether this result is due in part to a 'settling down' of the syllabus is difficult to prove explicitly, although with modular schemes the influence of early, perhaps less effective, modules is felt for longer as they may be included for certification up to four years after having been taken. This is very different from linear schemes where each year's assessment is self-contained.

The fixed part of the model is plotted in figure 8.15 below:

Figure 8.15 Fixed Effects from Combined Model

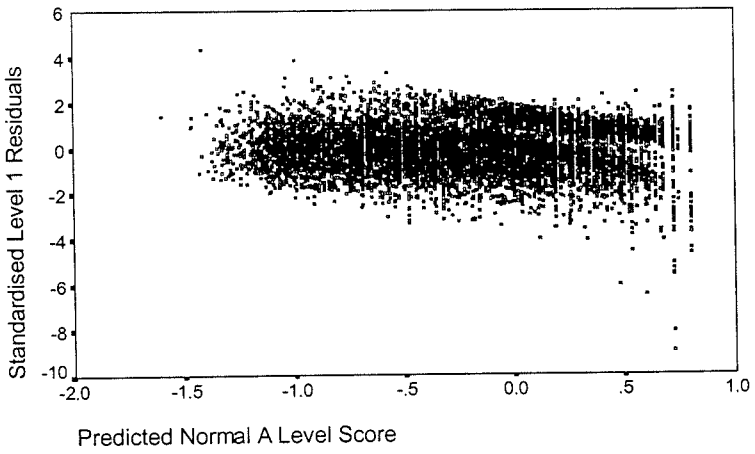


This graph is very similar in pattern to that of figures 8.11 although there are double the number of lines because of the addition of the year dummy variable, with again the gender effect outweighing that of the syllabus. There is clear parallelism in the

eight curves indicating that the predictions will be the same, to within a constant, of each other.

The standardised residual plot again holds no surprises.

Figure 8.16: Standardised Plot of Level 1 Residuals



The residual plot is reasonably well behaved and in line with previous findings with the same trend towards a ceiling effect. This indicates that the model is probably a good fit with the distribution of standardised residuals fairly even across the range of scores except at the very top end of the ability range.

The three level random part of the model gave the following as significant:

Random

Level 1

	e	γ_1	γ_2	γ_3
e (cons)	0.61 (0.02)	-0.10 (0.01)	-0.14 (0.01)	-
γ_1 (meangn)		-	0.03 (0.10)	0.06 (0.01)
γ_2 (syll)			-	-
γ_3 (sex)			-	-

Level 2

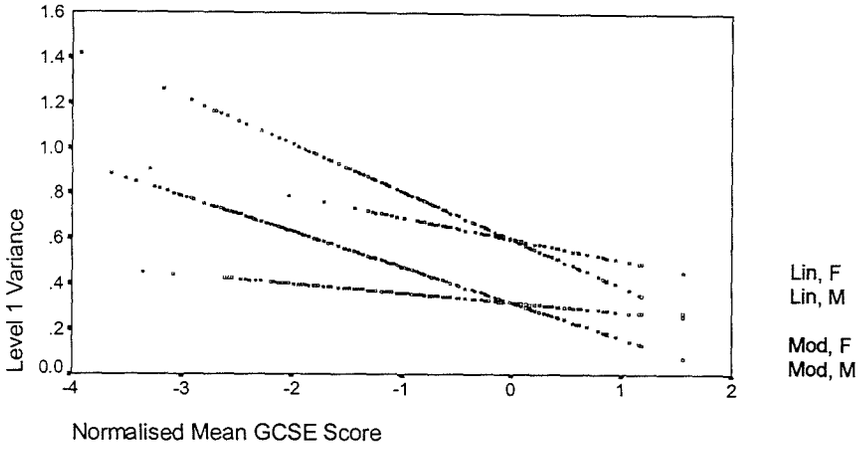
	v	δ_1	δ_2
v (cons)	0.04 (0.01)	-0.01 (0.01)	-0.06 (0.01)
δ_1(meangn)		0.01 (0.00)	0.01 (0.01)
δ_2(syll)			0.09 (0.05)

Level 3

	u	β_3
u(cons)	0.04 (0.01)	0.01 (0.01)
β_3(sex)		0.02 (0.01)

The level 1 variance plot is shown below:

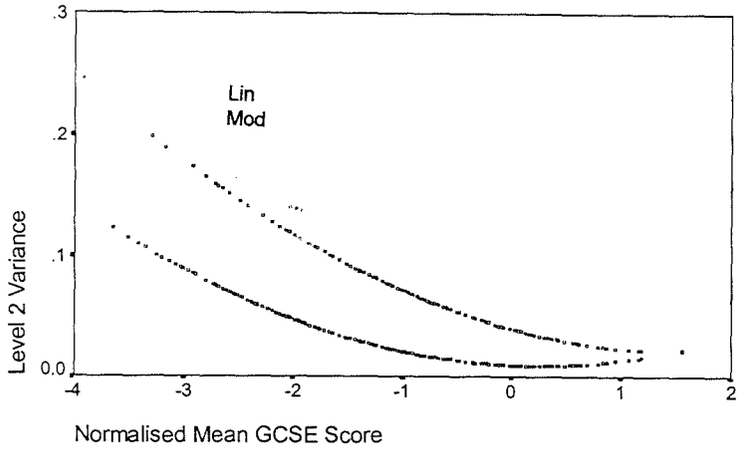
Figure 8.17: Level 1 Variance



The level 1 variance picture mirrors that found for the 1994 data, with both gender and syllabus random elements contributing to the total. The rise in variance with decreasing GCSE score is more marked for boys than for girls indicating an increasing unpredictability in A level scores with lower GCSE scores, and again the variance is greater for linear schemes. The increase of the level 1 variance with decreasing GCSE score indicates that at this end of the graph the normalised mean GCSE score is increasingly unreliable as a predictor.

Of most interest is the level 2 variance, i.e. the variance between years. This is shown below in figure 8.18. There are one or two points worth noting . Firstly the variance is, relatively, small, less than half a standard deviation in mean GCSE score. Secondly that there is an increase with decreasing mean GCSE score, with the variance for the linear schemes again greater than for modular schemes between years. This is an important point because the greater variability in one of the schemes between years (albeit small) indicates that what, comparatively, may have appeared as an inconsistency in the modular scheme, might have been due, in part, to a change in the parameters governing linear schemes. Clearly the inter-relationship between the schemes between years is complex and influenced not only by possible variations in grading standards but also by demographic fluctuations.

Figure 8.18: Level 2 Variance



The inter-centre, level 3, variance appears to be independent of mean GCSE score and syllabus, although there is a gender effect with the greater variance for the girls. Again the variance value calculated by the model is small, and the inference is that once inconsistencies between years have been allowed for, then any centre level effect is minimal. If this result is robust (and further analysis of between year data would confirm this hypothesis) then the implications are quite profound in as much as year-on-year differences could account for many of the perceived variations between syllabuses and these would be ephemeral in nature. Thus any result of a within year comparison may well be overturned the following year because of the dynamic nature of the examination system.

Such a result would also confirm many of the inter-Board comparability studies which, when repeated (unpublished) often fail to endorse a previous finding.

Discussion

There are numerous and complex reasons for the results presented here suggesting that differences between linear and modular schemes are far from being as unambiguous as other studies in this area (Tymms and Fitz-Gibbon, 1991; Taverner and Wright, 1995) have apparently found. The transformation to normalised A level grades and the normalised scores of the mean GCSE grades suggests that little significance can be attached to the differences in predictions at the candidate level, although at centre level, some effect can be detected - from analysis of the 1994 dataset. The same investigation, when repeated for the 1995 data, indicates a

reduction of such differences as were found. The greater representativeness of this dataset may be one reason and caution should be applied to analyses of this type when the incomplete data may be far from unbiased. The results from examination of the combined data set confirm that differences between syllabuses should indeed be viewed with some scepticism since between year differences suggest that variance from other sources, e.g. centres, may be exaggerated when considered from the perspective of a single year's results.

Additionally other comparability investigations have tended to group together all linear syllabuses and all modular syllabuses. Although no differences could be detected between the two linear syllabuses from the 1995 data, probably because of the very small numbers from one of the syllabuses, the assumption that any differences between linear syllabuses are insignificant compared with their differences from modular syllabuses is not endorsed by recent comparability studies (Quinlan, 1996). Therefore it is important that any analysis should be conducted at syllabus level, because each is distinct from its fellows, and not just in the assessment regime.

The gender effect is not particularly surprising given that girls tend to outperform males at GCSE, a result that is (currently) reversed at A level (Arnot, David & Wiener, 1996) and therefore the predictive capability of the mean GCSE scores is likely to be somewhat depressed for girls. That the predicted A level result for the best candidates is the same for linear boys and modular girls further detracts from any unequivocal finding.

What may be concluded however is that the level 1 and level 2 variances are great enough to lead to some inaccuracies in the predictions from the fixed effects part of the model. The one facet of interest is that, under the transformation, school level variances are not only gender related, but also depend upon the mean GCSE score. Because of the extent of the these variances, it cannot necessarily be concluded that either of the schemes is "more lenient" or easier than the other. However, it is a reasonable assumption that candidates may, given a certain level of ability, be more likely to show a higher level of attainment under a modular regime.

What does seem clear is that whatever differences were apparent in 1994, using the mean GCSE score as the predictor variable, were much reduced a year later. The

reduction in centre level variances suggests that a learning process may have been going on, but it may also be that the modular scheme had finally 'settled down' and become more integrated into the total A level provision. Additionally it is possible that standards in the 1995 linear syllabus were lower than those in 1994. The year-on-year differences may explain sufficient of the variation such that centre level differences are minimised.

There are detectable fixed effect differences which create some doubt about the underlying assumption of comparability of grading standards, but the indications are that the 1995 results suggest a closing of whatever gap may have been there in 1994. Of importance is the finding that it is the weaker candidates who seem to be gaining most by following a modular course of study, a finding which is consistent with those of previous chapters. The significance of between year variances, though small, should not be under-estimated since they should be instrumental in determining a policy which would allow a 'cooling off' period for new syllabuses. Reactionary changes are becoming the norm in the public examination system, arguably to its detriment. Year-on-year stability would seem to be of some importance, especially given the temporal nature of modern assessment methods. Such stability can only be achieved if growth is organic and not enforced.

CHAPTER 9

On to the Last

The focus of this research has been narrow. It has been concerned, first and foremost, with the anatomy of a form of assessment which, in recent years, has become more prevalent at A level, and will soon predominate in all subjects. We are concerned with what Broadfoot (1996) calls the "efficiency of assessment practices" and not with the wider issues involved with social engineering; but that does not mean that these should be entirely neglected. For example, it is instructive to consider why the reforms at 18+ took the form they did; what cognitive and pedagogical needs they satisfied; and what criticisms of the current provision they answered (see chapter 3).

Comparability, in the examination context, is itself an abstract concept. There have been many attempts to define precisely what it means (chapter 2) but many of these are untestable. If, however, we make the assumption implicit in the Cresswell social definition of comparability that grade standards are comparable because awarders are credited with the ability to take into account all the variabilities between examinations (in the same subject) when setting those standards, then we can investigate statistical comparability by considering how far equivalence stretches using grading information to test hypotheses. These variabilities and their differential effect on traditional linear and modular schemes of assessment are the subject of chapter 4, and are discussed in an attempt to address the question of why, if standards set by awarders are genuinely the same across syllabuses within a given subject area, there might appear to be differences in the grades awarded to similar cohorts of candidates. It is postulated that it is the very nature of the assessment which tends to enhance performances in the modular scheme under discussion rather than any decline in grading standards. The feedback which is possible within modular schemes and which allows an element of formative assessment even for written examinations and which is widely used by 75% of the cohort (see figure 5.4) is prime.

Variability within Modular Schemes

The three major modular concerns, or variabilities, which might compromise standards are those of resit and choice and, ultimately, coherence. These are the

focus of chapters 5 and 6. An important consideration is the use of standardised marks (or UMS scores) in the investigations. These are only valid measures if standards of grade setting are consistent. Whilst it is difficult to point to any positive evidence which directly addresses the issue of comparable grading standards, there is no negative evidence, and indeed the very high level of internal consistency found re-enforces the belief that the assumption of comparability of standards holds. There is evidence that modules sat terminally attract lower scores, but this is consistent with the gains accrued from resitting and the nature of mathematics and is not necessarily indicative of any variation in standard.

Of all the aspects of modular schemes, one that has given rise to the most controversy is that of resits - which appear rather radical in the A level context. As is sometimes the case with novel ideas, acceptance is less than immediate partly because of what may happen, not because of what does. The evidence presented here shows that candidates did, on average, resit once or twice, but that this would be for different modules. An average gain on perhaps two modules would be about 10 UMS points, or one sixth of a grade. Since most candidates gained an advantage through resitting (and there was no evidence that modules actually became easier or that gains were greater than could be expected had differences just been due to error) the conclusion must be that resitting allows candidates an opportunity to improve their attainment in those skills and knowledge areas tested by the retaken modules, and indeed the improvement continues with the number of resits taken. Whilst this sits uneasily with the dictum that there should be limited opportunities to retake any element of an examination within the period of study, it is also a vindication of the modular philosophy which is to improve a student's attainment through mechanisms of feedback, multiple opportunity and motivation. It is only possible to conjecture on the merits of improved motivation, but questionnaire and anecdotal evidence leave little doubt that the focusing of effort which results from in-course testing is generally felt to be beneficial. Worries that candidates would be forever re-sitting and gaining a grade or two as a result are not borne out by the evidence, and indeed, it might be said that the resit facility is used very conservatively by the majority of candidates. This leads to recommendation 1:

Within a modular scheme of assessment, resits should not be restricted.

The idea of choice within a scheme of assessment is not innovatory and exists as a possibility in many A levels where there are different options. However it is fairly new for mathematics where preference in conventional schemes is usually exercised through the medium of question choice within a paper. Since three modules are compulsory within the modular scheme investigated here, in practice it turns out that choice is usually made between a second statistics module, a second mechanics module or a coursework module. Detailed analysis of this revealed little of significance on the performance across modules and though there was a tendency of girls to choose statistics in preference to mechanics, with the reverse for boys, there was no gender performance differential. There was one obvious exception that of a module whose assessment scheme (now changed) was 100% coursework. It had been taken by the weaker candidates and there appeared to be significant differences between performances on this module and others. This finding might suggest that where there is choice, the scheme of assessment for each choice should be the same and hence recommendation 2:

Where there is choice, the methods of assessment within each non-compulsory module should be the same.

This is no different from conventional linear schemes where options may well differ in their assessment methods and this could lead to inequity. The internal consistency, one measure of reliability, of the examination was shown to be consistently high for each of the combination of modules.

Of equal interest is the manner in which the weightings of the different modules vary depending on the particular combination in which they are found. Whilst it was not the intention of this research to investigate the various methods and types of combination, it is true that changing the conversion methodologies may also affect the ability of a module to deliver its proper weighting. This is one area which may well prove fruitful for further research.

One of the criticisms levelled at modular approaches to the curriculum is that they may fragment the educational process. But that does not necessarily imply that the course, and this is certainly true within a subject, lacks coherence. In an attempt to allay these fears, all modular A levels are now required to include a 'synoptic

element'. Such an element is supposed to provide a link between all the modules - whether it is needed or not.

The research described here found that there was a considerable level of consistency which was firstly indicated by the high correlation between the modules, reinforced by the reliability finding, and this despite the fact that the modules have been distributed in time. However, the main vehicle for investigating more fully the consistency of the examination including the effect of seasonal, resit and gender variation, was a multi-variate, multi-level model. In the event it appeared that, to the model, the data was too linearly dependent to distinguish between modules. Whilst this is, in a sense, a negative result, there is the clear finding that coherence, as defined by module dependence, is sufficiently high for a sophisticated model to be unable to detect any functional difference. This suggests that once the contributions to the variance of resit and session are removed, there is very little additional information given by six modules over three. It also indicates that the error variance is small. There is, however a strong recommendation that:

The multi-variate multi-level analysis should be carried out on other modular schemes where differences between modules are more apparent.

A simpler formulation of the multi-variate model did allow investigations for each module to be conducted. The findings confirmed that resits did explain some variance, especially at centre level, and that, in general, modules are best sat within the usual two-year course of study. This probably means at the end (or shortly after) of the teaching unit for that module, but there is no data from which to confirm such an assertion, although clearly the majority of candidates are distributing their module sessions throughout the course. The results where session is a significant factor is almost certainly an indication that it is the weaker candidates who take the early modules terminally, rather than any difference in grading standards across sessions, although the latter cannot be entirely ruled out. More importantly, the individual module results show that, in general, the fixed effects predict lower module results for resit candidates. Since earlier analysis shows fairly clearly that candidates improve with resitting (although there may be error considerations, albeit small) we also now know that it is weaker candidates who use this facility.

Interestingly, gender within module comparisons is never a factor. The model also indicates considerable variation between centres in module results probably because of the greater potential for control over assessment that modularity gives to teachers.

Throughout the duration of this research, many more modular syllabuses have been introduced into schools, and a number of additional problems have been identified. That the MEI syllabus has managed to circum-navigate many of these is a tribute to the scheme's designers. The scheme has always been unashamedly modular, and there has never been a difficulty with candidates who choose to sit their modules in one sitting. They are still modular candidates, with all that implies. The use of design grade thresholds to which papers are written and a UMS conversion scheme which relates to it has also ensured that problems with conversions (mis-matched raw to UMS scales, very unequal, short or long grade bandwidths) are rarely found. However, the insistence of some schemes to retain a linear philosophy when grading some candidates, instead of considering them as a particular case of modular candidates and treating them as such, has led to inequities never experienced by the MEI scheme. In that sense the latter has proved somewhat unrepresentative and this leads to another recommendation:

More modular syllabuses should be investigated for internal coherence.

Although coherence is demonstrably high for MEI, it may not be for all other modular schemes for the reasons advanced above.

Variability of Demand between Syllabuses

The wider issue of comparability with linear schemes of assessment brings with it the central theme of fitness for purpose. Even within the limited contexts considered here, the simplicity of the research questions disguises considerable complexity. The first pivotal issue of the equivalence of grade distributions is, at best, näively stated since crude comparisons of distributions are worthless unless some measure of the ability of the candidature can also be included. So, too, is the question of the continuity of standards, which can only be considered by disentangling the constituents of what this type of equivalence entails. However, under the over-arching assumption of equivalence of grading standards, we need to consider the

variabilities in the domain of behaviour which need to be reconciled by awarders in making their judgements.

Central to any comparability study is the question of demand. Demand may be affected by a number of factors, subject matter and difficulty, question structure and context, scoring requirement, all of which must be taken into account in the setting of grade thresholds. Although there is slight difference in the perceived demand as specified by the one linear scheme and one modular, a close investigation of the two sets of papers revealed that syllabus coverage is greater in the modular scheme, although this is to some extent offset by a greater depth found in parts of the linear scheme. In this detailed analysis there was no evidence of construct under-representation which could be construed as threat to the validity of either scheme.

However, it is a truism that there is little point in asking questions, however penetrating, which the majority of candidates find inaccessible. Gratuitously reducing the discrimination of questions by making them too difficult or too easy helps no-one, and as the A level cohort continues to expand there is a clear duty to set question papers which discriminate effectively. The detailed analysis of question performance found that there was broad comparability between the questions set in the two schemes and that they elicited similar types of performance. It was at the extremes where differences could be detected - in the linear schemes there was a perceptible floor effect as some questions were answered very badly and hence failed to carry their correct weight in the paper totals. The reverse was true of modular papers in that there were questions which failed to carry their correct weight, but in this case it was because they proved too easy. Although these effects are largely immaterial since they would be offset by grading boundary decisions, it may explain why some public perception of the leniency of modular examinations arose. One aspect of the analysis does demand further attention and that is the link between the time taken to answer questions and the marks awarded to the question, given that the mark tariff is a measure of demand. Thus:

Further research should be carried out to determine the link between the time taken to answer questions, and question demand.

Whilst the analysis shows that, at the same grade, linear candidates need a lower percentage of marks, this is in line with floor effects and it can be concluded that there is little evidence for any disparity in grading standards. The assumption that candidates do, for the same grade, exhibit similar levels of performance has not been violated by the findings in chapter 7.

Another encouraging finding is that of a probable increase in reliability found in modular schemes. This is to be expected given the more homogeneous nature of individual modules although it probably does not impinge on the comparability issue explicitly.

Variability of Performance

Once there had been a scrutiny of the effect of variabilities in choice and demand, it was logical to then consider if, taking into account these effects, equivalent candidates, as defined by GCSE score, were gaining equivalent grades at A level. Multi-level modelling was used to set up a conditional model to investigate performance across syllabus and time. A pattern was established which showed that at the top end of the ability range, gender was at least as big a factor in determining performance as was syllabus. However, for weaker candidates there was clearly an advantage which accrued from taking the modular scheme. Again this is very much in line with expectation and anecdotal evidence gleaned from school questionnaires and with evidence from the previous work which suggested that it was weaker candidates who used the resit facility most and raised their performance level by perhaps 7 or 8 UMS per module. The half a grade syllabus gain could therefore be explained by the resitting of, perhaps, four modules. This suggests that grading standards could still be comparable between the two schemes, but that the availability of the resit option, especially to weaker candidates, whilst not compromising judgemental comparability, enhanced performances. In short, the improvements in the value–added by the modular scheme from the GCSE base have been shown to be a legitimate result of the different constructs of the syllabus. It would therefore be a construct-relevant variance and would be no threat to the validity of the modular syllabus as an evaluation tool.

One of the advantages of using multi-level modelling is the added insight into predicted performance given by the variance analysis. These variances are not only GCSE dependent, they are also affected by the influence of gender. The extent of the variance at centre level is sufficient that differences in fixed effects which result from the two syllabuses may not be very significant.

The final finding of chapter 8, using a three level model, is that whatever diversity may have been apparent in 1994, it was reduced a year later. One explanation is based on the finding of reduced centre level variance when year-on-year level variation is taken into account. These level 2 yearly variances, which increase with decreasing GCSE score, indicate that any difference in predicted performance as a result of syllabus effects should be treated with caution, especially at the lower end of the ability range where the year-on-year variation is greatest. We have therefore a further recommendation in order to ensure that the year-on-year finding is not an isolated one that:

More year-on-year studies should be undertaken both between the same syllabuses, and similar syllabuses to determine whether these differences are found elsewhere.

Thus, although there are differences in the fixed part of the model associated with modular/linear effects, and these may well indicate some leaning towards leniency in modular schemes, the variances serve to make these differences far from unequivocal. Certainly there is little evidence that candidates are gaining by whole grades under modular assessment regimes.

What is of some interest is the gender effect found in performance. Whilst these may show up some bias in the schemes of A level assessment considered, since they do not appear elsewhere in the analysis, it is far more likely that this difference is as a result of differential GCSE performance than of achievement at A level. This needs to be replicated before it can be given as an unequivocal finding so:

Further analysis should be undertaken to investigate gender difference in performance at GCSE, based on A level performance.

All in all, the various facets of the research reported here combine to paint a picture of a scheme of assessment that in execution is very little different from conventional schemes. It is also a reasonable deduction that weaker candidates do perform somewhat better under such a regime because of the relaxation of the all or nothing approach. But it is emphasised that this improvement is indeed because candidates have gained from repeated testing, and that for some this may be educationally desirable, though perhaps counter-productive for some of the high-fliers. There is no evidence of any lowering of standard in either demand or performance and some evidence that both reliability and validity may have been improved.

Rationale and Generalisability

The rationale for this research was founded on the need to consider, in detail, the anatomy of a new approach to assessment at A level. When it was started in 1994, there were few modular schemes of any size and only one in a 'main stream' subject. Much uncertainty surrounded the question of the acceptance of the modular assessment methodology, by centres, teachers, regulatory authority and the public, and as has already been suggested, few of the latter modular syllabuses appear to have been as well constructed as the MEI syllabus.

Bennett and Dunne (1994) state that "learning involves the construction of knowledge through experience". The focusing role of testing has always been accepted, and this seems to have influenced those who are the decision makers. Certainly the considerable rise in the popularity, allied to their availability (which is not unrelated) of modular schemes since the start of this research has been unexpected, and such assessments are now available in most subjects. That does not necessarily mean that all subjects are best suited to modularity, rather that there may be a band wagon which is rolling and students want to be part of it.

Therefore one consideration must centre on how much of the findings reported here can be extended to other subjects. Integral assessment will be as beneficial in any subject, but there are maturity effects which may outweigh such benefits. What little evidence there is suggests that patterns of behaviour are similar for science subjects, that candidates do take some modules early, and also resit with similar frequencies. The educational experience is thus likely to be the same.

The same cannot be said for the more discursive subjects such as English. Here skills are strongly allied to maturity and extend across all modules. This factor alone suggests that whilst there may be benefits to be derived from modular schemes, these are likely to be limited. One result which may be applicable is the inadvisability of allowing different assessment methods within options. For example, it would seem from the results here (e.g. data on modules 19 and 21, see chapter 5) that to allow modules with, say, 100% coursework to be equated to a wholly examined option, would be to introduce an added source of inconsistency (a finding not unknown in linear GCE examinations where options involving different assessment methods are allowed). However, it is not possible to disentangle the nature of the content from the nature of the assessment method and the cohorts for these particular modules are far from representative. Again such data as exist point to the majority of modular English candidates taking all their modules in one session, terminally (SCAA, 1996).

Question structure has also had an important part to play in the MEI mathematics questions as witnessed by the ceiling effect. It is a feature of modular testing that the examinations are shorter, contain less question choice, if any, and are far more focused than linear schemes. All these features may enhance performance and attainment illegitimately if not accounted for in grading standards, and it is something which should be investigated on a wider subject base before generalisations can be made with any authority. Against this is the argument that the repetitive testing that goes on in the revision sessions associated with linear schemes may be just as enabling as any form of question structuring. However, in this research, there is no positive evidence that awarders have not taken these variabilities into account when setting grade boundaries.

However, one of the biggest problems with generalisability is in the way centres choose both to teach and enter candidates for assessment. Whilst all modular schemes subscribe to the same philosophy, how it is applied differs both between subjects and teachers so much so that the traditional pedagogy may be unaffected. For MEI and those who take it the pattern of both teaching and assessment is very clear and where similar patterns exist in other schemes, the findings here may well apply.

There may also be a question mark over the universality of application of the UMS conversions used in the aggregation process. MEI uses what is already a 'non-standard' application which is entirely appropriate given the structure and rationale behind the scheme where paper design thresholds are matched with UMS conversions. But it is recognised that it would not be suitable for all subjects, and the attempt to impose a universal conversion would not be recommended.

Why Modular?

Almost since its inception, the A level regime has attracted criticism for its narrowness, early specialisation, its inability to respond to change; but although there were some desultory attempts at innovation, there appeared to be little will, probably because the universities themselves were not keen to see any dilution of the perceived rigour of A levels. This would be the expected outcome of any reform. Despite comprehensivisation, what Hargreaves (1982) calls "the hegemony of the grammar school curriculum" predominated. But the population of A level takers was growing throughout the 70s, and with it their expectations. Although the number of different subjects increased with demand, traditional syllabuses still continued using traditional assessment methods. They were seen as 'hard' by many candidates who would avoid them if possible. Although the subject content could not change, the manner of its assessment could.

If there is any agreement between today's writers on assessment it is that it is multi-functional and multi-faceted. Evaluation has progressed some way from its original purpose of selection by means of a single session written examination. Harlen, Gipps, Broadfoot and Nuttall (1994) identify four roles of educational assessment: formative; summative; certification; evaluative or quality control. Apart from the formative aspect, any public examination can claim to fulfil, at least in part, three of the four roles. Modularity allows the fourth (as defined in chapter 4 as any assessment which allows testing *within* a course of study) to be met.

Of greater import is the meeting of more of the key principles of assessment outlined in the same paper. These are that assessment must be used as a continuous part of the teaching-learning process, and that assessment should "serve the purpose of improving learning by exerting a positive force on the curriculum at all levels". It is this latter point which is contentious. If, by improving

learning, attainment is also enhanced then this is a function of the assessment tool, not of a lowering of standards. It is a legitimate variability - it is also immensely difficult to prove.

It has always been understood by education professionals that at another time, on another day, with a different teacher, any candidate is capable of producing a different measure of attainment, though it is a perception not often exhibited by the public at large. I would postulate that there is a range within which it is expected a given candidate's performance will usually lie. Whatever variabilities are considered, the measured achievement of an individual will generally fall within this range. Additionally the range is very unlikely to be more than two grades, but is quite possibly one and will depend on the coherence of the examination. With MEI, for example, the range is at the low end because of the high correlation and internal consistency of the assessment (see chapter 5). If all the variabilities are positive, then the grade achieved will be at the top of this range, if they are all negative then it will lie at the bottom. Modular assessment will, in general, have a constant positive effect.

As the part that public examinations play in the lives of school age children has become more dominant, legitimate sources of criticism, namely the stakeholders, have become more varied and more insistent. It is important, therefore, to address the question of the effect of such assessments and the broader policy issues involved (Broadfoot, 1996). It is fundamental to any discussion on the raison d'être of modularity to consider, not just the means but also the ends (ibid.). Even though the means can be demonstrated to be equivalent, the ends may not be so.

Theodossin (1986) attributes much of the rise in modular schemes to a "power shift" which has seen a change from "provider-centred" control to "client-centred", describing modularisation as "customisation", and explains much of the shift in terms of demography - a decreasing 18+ population, and increasing number of A level certifications. Though this may be true, it is probably only part of the story. With the increasing percentage of 18 year olds taking A levels (consequent upon changes in policy at 16+), there has been a rise in the number of pressure groups, including the government, who have a stake in seeing that standards are maintained. With league tables such interests also entered the public domain.

In the days of the traditional A level, the primary, perhaps only, stakeholder was the candidate, since the universities would consider themselves grouped with the boards as providers. (Employers had a very low profile, given that the vocational bodies were empowered to provide directly for their requirements). They, the pupils, had no means to make changes or power to enforce them. As media interest increased, the government also became stakeholders, their educational policy was judged by the results obtained. However, they were in a cleft stick - what to one observer is rising standards - 'more candidates pass', to another is falling standards - 'exams are becoming too easy'. The publication of league tables meant that schools vied with each other to attract the best pupils, to get the best results, to get the best pupils. This gave parents real power, they could refuse to enter their children for schools which were perceived to be failing on the basis of examination results. And, of course, teachers were pitted against each other in order to improve results.

It is not surprising that in this maelstrom of interests there should arise some means by which the pressure could be alleviated. If candidates could be offered genuine choice, including that of withdrawing before that final examination, if teachers could apportion modules so they were no longer competing, most of all if the motivational effects of distributed assessment could be seen to be positive, then such a scheme was likely to be welcomed. Overall, results tended to improve for schools opting for modular syllabuses, although it was partly a false improvement because a number of candidates would have chosen to withdraw. On the whole though, there were positive effects, especially amongst weaker candidates for those schools who were prepared to use the scheme to their best advantage, as shown earlier in the enhancement of results due to resits. The facility to maintain standards while simultaneously improving performances has been one of the holy grails of assessment, and modularity appeared to be one way of achieving just that.

There was also an additional player on the scene. For many years, schools had been prohibited from offering vocational courses. But, with the institution of the NCVQ in 1993, these awards were no longer the exclusive province of the FE colleges. But they were considered inferior to the academic A level. One means of changing the public perception of vocational qualifications was to provide a 'mixed' qualification, part academic, part vocational. Such was perfectly possible with modularity, and it is no accident that the original champions of modular schemes

were the vocational bodies. They were thus additional, and powerful, stakeholders, and consideration of the pressure that such groups could bring does explain one force for change, and points the direction that change was likely to take.

The recurring strands of curriculum, pedagogy and evaluation, which together create today's educational experience (Bernstein in Broadfoot, 1996; Murphy and Torrance, 1988) are all influenced by the choice of examination syllabus. But the gradual changes which have been wrought in each of these strands have led, almost inevitably, to the need for a more flexible approach to examining. It has been suggested that examinations are more reactive than radical, and this is so. It is however possible to see within what appears to be radical, merely a reaction to social and pedagogical reforms. Broadfoot (1996) sums up the changes in "the information society" thus:

In place of the emphasis on knowledge acquisition and the associated emphasis on didactic teaching approaches, the beginnings of a quite different emphasis on education as the facilitation of skill acquisition is already apparent. (p63)

In order to include evaluation of, if not skills based, certainly skills inclusive, learning, mixed assessment methodologies are necessary. These can be accommodated in linear schemes, but there is, within, and as a result of, these changes, a tendency towards less deterministic assessment. Schools and candidates have come to expect a distributed policy towards evaluation, not only because of the GCSE experience but also because of the increasing heterogeneity of the population taking A levels.

Based on the GCSE experience, students are likely to expect more from a two year A level course than the traditional approach would give them. There has been a realisation that even public examinations can have a formative and diagnostic role and actually add to the education experience. Karl Popper (1972) has stated "we are fallible, and prone to error; but we can learn from our mistakes".

Part of the pedagogical process is the business of 'learning from our mistakes', and this is especially true in mathematics. One practices sums in order to make mistakes in order not to repeat them! But there is a fundamental truth regarding assessment hidden in Popper's words. This is the effect that assessment has. If it takes place at

the end of the course, it cannot affect any other assessment of achievement of that course. However if it occurs, in part, within the course itself, it can most certainly affect attainment in any subsequent assessment. The experience of sitting the examination can, by itself, have beneficial effects. There are, not only, positive effects which are explicit in resits, but they are also, almost certainly, implicit in any future module examination. In many cases when a candidate resits a module there is little further teaching, but the evidence is clear that candidates who resit, do improve their performances (see chapter 5). Even so, these candidates are also weaker than those who do not resit (see chapter 6). There must be a causal relationship between the two sets of results, and it would be difficult to deny the possibility, probability, that the stronger candidates (who need not resit) also benefit from taking modules within a course. They learn to focus on the assessment, to work consistently throughout the course, to learn from their experiences.

The key point is that assessment influences assessment. In quantum theory it is impossible to know both the position and momentum of a particle accurately. Any attempt to measure the one with any accuracy will immediately affect the other. My thesis is that the same is true of modular assessments, but it is one that can only be conjectured. No one candidate can both take and not take early modules. There is therefore the expectation that candidates in modular schemes who take advantage of the flexibility of the examination sessions will improve their performances simply because they have taken part of the assessment early. The belief that candidates may be performing to a higher standard under the benign influence of modular assessment would thus have some rationale. And it has nothing to do with easier, or more difficult, or even differing standards. It is a function of the assessment form, a form which we might term integral assessment i.e. it is a defining element of a course of learning, not separate from it. The means may appear identical, the ends are not. We have therefore the proposition that:

Integral assessment enhances achievement.

This theory also leads to the conclusion that modular schemes are less deterministic than linear schemes, a theory only partly supported from the multi-level analysis. In the linear scheme there is no temporal choice, all assessments are equal (or supposed to be), in the former the facility to take some modules within the course will lead to a different outcome from one that would have obtained had such a

choice had not been exercised. Modularity does not make examinations easier per se. I believe that it is the recognition of the educational benefits of within course assessment which is one of the reasons for the increasing popularity of modular examinations. It leads to another proposition that:

Modular schemes are less deterministic that linear schemes

and this post-modernist, educationist view might be at the heart of much of the unease about modular assessments. It implies that two candidates of equal aptitude might indeed gain different grades because they pursue separate testing programmes - even though the grading standards of the modules taken are equivalent.

It is not only Cockcroft who could see benefits to the pedagogical implications of modular approaches to the curriculum. Seymour Papert, the great guru of microworlds, drew numerous comparisons between the logic of learning how to use a computer and wider epistemologies. He has interesting things to say about 'debugging' a program, but more importantly, when extolling the virtues of the modularisation of knowledge, he says:

When knowledge can be broken up into "mind-size bites," it is more communicable, more assimilable (sic), more simply constructable..... (p171)

As a validation for the pedagogy of the modular curriculum it could hardly be bettered especially, as Papert acknowledges, since it builds on the work of Piaget. However, it is important that such a concept is not taken too far and that in the attempt to feed manageable, bite sized pieces the whole becomes too atomised. It is a warning that Vygotsky would have endorsed since he espoused a holistic approach. He proposed that the whole be partitioned into elements, each of which contained all the basic characteristics of the whole (q.v. Moll, 1990). Extending that to this particular research, it is clear that a Vygotskian procedure would insist that not only should each module be self-contained but also replicate the same assessment structure.

Other reasons for the growth of modular schemes concerns the nature of the populations involved. Firstly there is the overall increase in sixth form population.

Whilst some of this may be from the most able, there has to be an assumption that much of the expansion has had the effect of including many of the less able (q.v. SC Working Paper 45, 1972). There is thus little point in setting an examination which is not only inaccessible, but for which the numbers passing remains stable year-on-year. Examinations are a method of categorising the cohort and it is an inescapable fact that failing the majority would signally fail in that categorisation. Only with an entirely stable syllabus entry (at least in ability) can standards be strictly maintained.

Siegel (1988), expounding on the merits (or otherwise) of minimum competency testing also declaims the arbitrariness of standard setting referencing Popham's distinction between "capricious" and "judgemental". Although there are clear guidelines for the setting of grade boundaries judgements cannot be absolute in the way that a measurement can. Statistical indicators are used increasingly, but again are only indicators because so many of the variables are unknown. The most important of these is the composition of the examination population, and this poses a conundrum for module awarders where module populations are not constant either in numbers or ability.

The needs of the populations taking mathematics have also become more diverse. No longer do those taking mathematics necessarily wish to study science. The rise in proportions of candidates taking mixed A levels (i.e. arts and sciences) from 10% to over 30%, plus the needs of other subject areas (e.g. psychology or economics) has meant that no longer is mathematics the province of the most technically able. The subject curriculum must reflect these differing requirements, and one way is to offer a modular scheme where there are a number of options encompassing these various needs (see chapter 3). Again this is a requirement foreseen by Cockcroft (1982).

Validity

At the heart of this research is the question of validity. If modular and linear schemes are valid, then it is suggested they are also comparable. In the cases investigated here, because the potential domains of assessment (as specified by the syllabuses) are virtually the same, although there is some exchange of breadth for depth (see chapter 7), the detailed analysis of the questions and subject matter of the two schemes indicates a high level of similarity. There is clear evidence that they are measuring the same things (chapter 7) and cannot therefore be said to exhibit construct under-representation. Additionally, though there is evidence of enhanced performance, this has been shown to equate to a legitimate difference in construct. These, though simplistic, indicate a measure of construct validity, which, according to Messick, is the overarching form which subsumes all others. If, too, the 'social' definition of comparability is seen as a property of the equivalence of use to which both schemes are put then this brings us full circle to the social implications of the uses of them, and indeed the definition of comparability coined by Cresswell (1996). I would suggest that as far as these are concerned, the imposition of the constraints imposed by SCAA restricts the generalisation of the assessments and certainly the contexts in which they are valid. However the potential for wider application exists for a modular scheme far more than for a linear one. Modular pedagogy will also have an effect which differs from that of linear schemes - the emphasis on self-discipline, continuous effort, different assessment forms, formative and diagnostic evaluative assessment constitute an unquantifiable different learning experience. Extension of these effects to alternative social contexts is desirable, but so ephemeral that one can only speculate as to the outcome.

In the End

Since the start of the research presented here there have been a number of changes in the A level provision, and new suites of syllabuses are waiting to be introduced. These are modular but are more constrained than at present especially regarding restrictions on resits. It is probable that the current modular schemes will, in time, be seen as a stepping stone, a trial for the future: their very popularity indicates that such schemes are successful, even though the administrative burden they place on schools is considerable. The new wave of syllabuses almost entirely

discounts the linear ethos, and, despite some reservations, modularity is here to stay.

If we can believe that grading standards are equal, then there is little in this research which would deny that assertion. However, there are variabilities which might lead to more stringent grading decisions on one scheme, or module, than another, although there is no evidence that this is the case. Where apparent discrepancies do occur, especially where candidates are deemed equivalent on the basis of prior achievement, they can often be explained by advantages conferred by the modular scheme. To summarise on the basis of this research, modular schemes are coherent and their flexibility is used most by weaker candidates. Resits do enable such candidates to enhance their scores, but within reasonable tolerances. There is evidence that modular schemes may be more reliable than are linear syllabuses. Season and resit effects are only significant in some, early taken modules and it is the weaker candidates who take advantage of the resit facility and although they gain, they are still the under-performers. There is little to choose in learning content and hard questions on linear papers and more accessible questions on modular papers are offset by the need for higher percentages of marks to be scored to obtain grades on the modules. One possible explanation for the perceived leniency of modular schemes and difficulty of linear schemes is the clear ceiling effect of some questions in the former and floor effect in the latter. This does not imply any difference in grading standards. Further evidence based on prior attainment indicates, again, that it is usually the weaker candidates, who are gaining most from the modular scheme of assessment although year-on-year comparisons compound doubt about clear evidence of a difference in grading standards. The circumstantial evidence is therefore that assessment can and does play an important part in determining attainment, especially for the lower achievers, but that grading standards are not necessarily compromised by the change in assessment patterns.

One conclusion is clear, and it is negative. There is no evidence that modular schemes expect less of their candidates than do linear schemes. Nor is there any evidence which points unequivocally to differences in grading standards within or between schemes of assessment. That which does is susceptible to other explanations than a violation of the prime assumption of this research. There are

undoubtedly a number of potentially illegitimate variablities, but there is no evidence that these have not been accounted for in the grading process.

There is one further aspect which must be considered. Neither linear nor modular schemes are 'complete'. There are a number of forms of assessment (even within a subject) which are not included in public examinations. For example, there is no attempt at setting 'mastery levels', no pre-testing of examination questions, no hurdles are set (although there is a minimum UMS score of 5 for the MEI scheme), not all schemes include coursework, multiple choice and so on. That the inclusion of other forms of assessment will have pedagogical and cognitive implications is a truism, but whereas qualitative judgements as to their effects can be made, quantitative differences, should they exist, can only be estimated. Increasing complexity will not necessarily enhance applicability or reliability.

In the end, belief in the comparability of assessment schemes is an act of faith, as articulated in the Cresswell social definition. None is complete, even in the modular context where flexibility is paramount. We regularly subject evaluations to processes which are probably indefensible. But the remarkable truth is that it does not seem to matter. The great mathematician Kurt Gödel in his work on number theory postulated in 1932 that there is no input to an algorithm whose output can tell you whether that input is true. In a very much more prosaic formulation, one might also conjecture that there is no output from an assessment which will tell you whether the assessment instruments, taken together, are comparable with any other assessment instruments.

But, now that you've stated the whole of your case,
More debate would be simply absurd.

Lewis Carroll, 1876.

Anomalies, UMS and the Regression Allowance

Aggregation issues are a source of constant debate within the public examination area. A good description of the more common methods is given in Thomson (1992). Many of these give rise to one of two types of anomaly:

Type I - Two candidates with the same grade profile receiving different subject grades.

e.g. abbd = B
abbd = C

A special case of this is a candidate who obtains the same grade for all components, but obtains a different subject grade

e.g. bbbb=A

Conventional grading as applied to linear subjects gives rise to instances of subject grade being different from the component grades, even when the latter are the same. This is because of the use of *weighted percentage* aggregation when allowance is made for the clustering of candidates' scores when aggregated over a number of components. This clustering results in a tendency for aggregate scores to be closer to the mean than found on individual components and is generally known as *regression to the mean.*

Type II - Two candidates with a different profile obtaining the same grade

e.g. abbc = B
aabb = B

Different methods of aggregation give rise to different instances of these errors. Unless there is a very crude system of assigning a point to a grade, all methods will result in at least some type II anomalies, and many to type I. One of the reasons for the choice of

uniform marks for aggregating modular schemes is that the instances of anomalies are reduced (Thomson, ibid).

UMS

A description of the conversion to UMS in the MEI structured mathematics scheme is detailed at the end of chapter 4, but there are a number of different schemes in operation, most of which are based on a 600 mark total. For an evenly balanced scheme of six, equally weighted modules, each module attracts a maximum score of 100 UMS marks after conversion, with 80 for an A, 70 for B and so on. This gives twice the A range than is used for MEI, but the other pass grades all have the raw grade range mapped on to 10 marks. If the modules are not equally weighted or six in number, or both, the UMS for each module is usually calculated to be in the proportion of that module of 600, with the boundaries set accordingly, so that in all such cases there will still be greater compensation for an A than any other grade.

One of the problems that has occurred in other modular schemes is a reduction in the expected number of high grades. Part of this is due to the lack of consideration of the effect of UMS conversions when mark schemes are devised and often much of the compensating power of a 20 mark UMS A range is lost because raw mark ranges are too long, or are not fully utilised. This has not been a problem with MEI because design thresholds ensure that there is a known matching (within the tolerance of the variability of paper demand) of grade bandwidths to UMS and predictable conversion effects. The other reason is the loss of a regression allowance.

One of the criticisms which attaches to the UMS method of aggregation is its invariance. Syllabus (and module) boundaries are pre-defined and thus not open to 'statistical and technical' adjustment *post hoc* such as may be found with linear schemes. If such variation year-on-year were allowed, then this could give rise to a third type of error. Candidates with the same uniform mark total could be getting different syllabus grades from year to year. Since raw grade boundaries are set to allow for differences in demand, the point about the UMS conversion is that this differential has been allowed for. Looked at from raw mark viewpoint, if syllabus grade boundaries are allowed to fluctuate, then the relationship of raw module boundaries to that final total will vary. Even calculating a regression allowance of UMS scores would

lead to year-on-year anomalies because candidates on what were ostensibly equivalent marks could achieve different grades purely because of the company they keep even though much of their assessment might be common.

The Indicators

Syllabus grades are an artefact of the constituent component grades. The GCE SCAA Code of Practice lays down the rules of aggregation for component grade threshold marks in order to arrive at a syllabus threshold mark for a given grade. For linear schemes of assessment this is achieved by the calculation of two indicators, 1 and 2, which then define a range within which the boundary must lie. It is usually the lower of the two indicators.

Leaving aside the question of whether aggregating raw scores is meaningful, indicator 1 is simply the addition of the component raw marks, suitably weighted. Indicator 2 is the mark gained by the weighted aggregate of the percentages achieving the grade on each component. The rationale for their use is based on the distributional characteristics of aggregate scores. If all components were error free and perfectly correlated the two indicators would coincide. They do not because these perfect conditions never prevail. Whilst it is desirable to reduce error as far as possible, it is arguable whether perfect correlations between components is equally so since if each component measured the same trait, as would be evidenced by high correlations, then only one would be required in order to rank and grade candidates. There are also sound educational reasons for setting tests on the whole of the contents of a course even if highly correlated.

Each of these factors affects the 'ideal' distribution of final marks (see Good and Cresswell, 1988). Error terms will always increase the standard deviation of this distribution (Guilford and Fruchter, 1986). Conversely, low component correlations will lead to bunching of the aggregate scores. This is because the lower the correlations the more ways there will be to get an aggregate mark and this will cause bunching or 'regression to the mean'. For example, in an equally weighted, two, equally discriminating, component examination, an aggregate score of 50 can be gained by candidates scoring 20 and 30, or 30 and 20, or 25 and 25 and so on. However this would only be possible if the rank order of candidates was different for different

components, i.e. the correlations were low. The lower the correlations, the greater the tendency of distributions to bunch about the mean score i.e. there is a reduction in the standard deviation from the ideal distribution. Equally, as the number of components to be aggregated is increased, so too are the number of ways of achieving the aggregate score and the greater the bunching.

In practice, the regression effect is always much greater than any error effect and this means that at the top end of the distribution, fewer candidates will achieve a high given grade boundary than would be indicated by the percentages gaining the grade at component level. Indicator 2, which essentially determines the maximum (above the mean, minimum below it) percentage which could obtain the grade and translates this into a mark, will produce a lower boundary mark than indicator 1, a difference that is reversed at the bottom end of the distribution.

Where correlations are low, the differences between the two indicators is high, and the allowance gained by the use of indicator 2 quite substantial. In modular schemes there is no allowance for indicator 2 and this is potentially a disadvantage to high achieving candidates. At the bottom end of the grade range, indicator 1 is the indicator of choice and there is no effective difference in the attainment required to gain a low grade between linear and modular schemes except those caused by conversion factors.

With the UMS system of aggregation it would be impossible to obtain a different subject grade from equal component grades, and instances of other types of anomaly are less prevalent than with other methods. However, the lack of regression allowance suggests that it is inherently more difficult to obtain a high overall grade in a modular scheme, especially when modules correlate badly.

The two indicators can, in practice, only be applied to linear schemes. There are three reasons for this:

(i) the population for each module is different. Therefore a percentage reaching a particular grade in a given module is not indicative of the percentage of the certification cohort reaching that grade.

(ii) while the same module may, in essence, appear very similar year on year, it is possible that its correlation with any other module (either at the same or different session) will not be constant. The rationale for making an allowance for regression is based on constant differences in correlation between components. The variations both in time and modules mean that there would also have to be variations in the regression allowance.

(iii) candidates taking at least some of their assessment at the same time may have to achieve different scores to obtain a given grade. The relationship of module scores to the final total will vary in unpredictable ways which will not be fair to some candidates.

The effect of the loss of the regression allowance can be calculated for the MEI examination (in terms of UMS scores). Although it is possible to ensure that the population for each module combination is the same, there will be slight differences in correlations over time which are disguised by the UMS conversion process which only accounts for differences in paper difficulty over time.

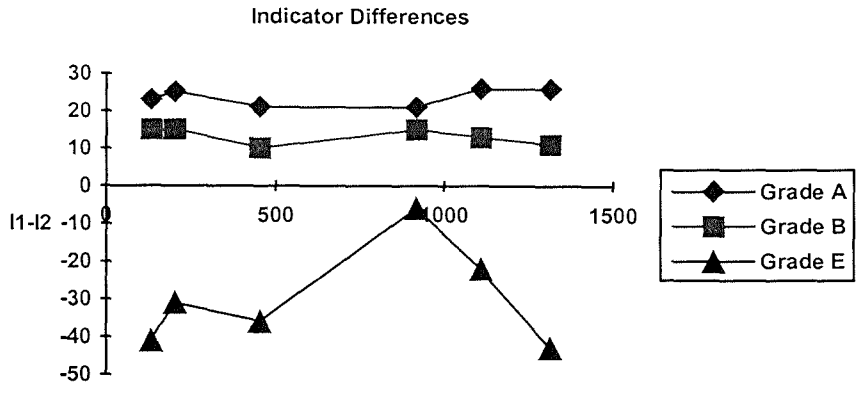
The six most popular combinations of modules have been taken and the potential regression allowance calculated. All candidates take modules 1, 2 and 3 with the other combinations shown in the table below:

Table 1: Module Combinations

Module Combination	Number of Candidates
13, 14, 19	134
7, 13, 19	203
7, 8, 9	453
13, 14, 15	919
7, 13, 14	1109
7, 8, 13	1316

The regression allowance (I1-I2) has been calculated for each combination and plotted on figure A1 below, with the abscissa indicating the number of candidates in each combination.

Figure A1: The Variation in Indicator Differences with Grade and Entry



In each case I1-I2 is negative at grade E, positive at A and B. There is also consistency across modules at each grade. Although there is some variation in the numbers taking each combination, this appears to have little effect on the regression allowance.

On average therefore, it might be expected that, if regression were to be a factor, at A it would be some 4% (23.7 UMS marks) easier and at B about 2% (or 13.2 UMS marks) to obtain the grade. This would result in an increase of 8.8% of candidates getting an A or a 4.7% increase at B and a proliferation of type I anomalies. To make such an allowance might also be seen to be against the spirit of modularity because correlation between modules would become a deterministic influence in the calculation of the grade boundary, though as has been shown, for mathematics at least it would be possible to use a constant allowance for all combinations of modules without introducing another source of inequity.

Conventional linear mathematics schemes which have two written components covering the same subject areas are highly correlated and the regression allowance is small (and in some cases zero). That there is a noticeable regression effect in the modular scheme reflects the mixed assessment methodologies and subject content of each module, both of which will tend to reduce the cross-module correlations.

Drawing the Line

It has long been argued (Angoff, 1984; Thorndike and Hagen, 1964) that where examinations are concerned, raw score scales are of little value. Whilst they may be used to rank order those candidates who have taken a particular examination at a particular time, as, for example, an entrance examination, they are often an unfair measure when taken in a wider context. The raw score has little inherent meaning of its own, and cannot (or should not) be compared with other raw scores from other examinations, because there is no common point of reference. In particular this is true of separate tests within the same syllabus where it is common to add together raw marks on the assumption that they have the same value.

Whilst Thorndike and Hagen suggest that only by comparison with a reference or norm group can a raw score have meaning, Angoff goes somewhat further in proposing that there are really three elements to the operation of ascribing value to test scores; namely that they need to be scaled, norm referenced and equated (or calibrated). The grading process in British GCE and GCSE examinations encompasses all three within a single awarding meeting.

Scaling (equating distributions) tends to be a 'by eye' process based on the mark distributions which may be brought to the meeting at various stages of the decision making process. If norm-referencing is taken to mean that marks are defined 'in terms of the performance of a representative group of individuals' and equating to be the conversion of scores from different components/modules to a common scale, then awarding a grade may be analogous to these processes. One essential difference between an awarding meeting and the calibration of other tests is that norm-referencing is carried out by experienced awarders exercising their judgements as to how a 'representative group' would have performed on a given component. Without extensive pre-testing it would be difficult, in practice, to carry out such calibration without reference to judgement, since the ever-changing nature of the entry population (especially for modular examinations) makes appeal to statistical methods alone less than foolproof.

However, norm-referencing in the context of grading in the UK is usually taken to mean the allocating of fixed percentages to particular grades (French et al, 1987). This almost

certainly leads to a different grade distribution from that which would be expected should marks be directly related to a representative, or norm, group and under this regime there can be little suggestion of equality of grades from different tests other than that bestowed by percentile equivalence.

There are, therefore, three major, interrelated problems which attach to the grading process. Firstly are the difficulties inherent in converting from one type of scale i.e. a raw score scale (which is rarely strictly interval despite appearances), to another, apparently simpler, ordinal grade scale. The second is determining where, on the raw mark scale, the line between categories, or grades, should be drawn i.e. the calibration of the marks. Finally there is the problem of aggregating marks so they fairly reflect the achievements of the candidates in the separate components.

Most public examinations at 16+ and 18+ are an aggregate of several assessments. The correlation between these different elements is never perfect and can sometimes be quite low. Even when correlations are high, the characteristics of the mark frequency distributions can be dissimilar. Aggregation is the simple process of adding a number of (possibly weighted) raw scores with an additional *post hoc* adjustment for regression if necessary, but because the raw scores for different components are almost always differentially distributed, like is not being added to like. Assuming that there is no adjustment for regression, which is certainly the case for modular examinations, aggregation reduces to the addition of component scores and it will be used in that sense here.

The current, compensatory method of obtaining a linear syllabus grade by the addition of component raw scores makes no concession to the differences in raw score value (as defined by component grades) between components and thus may be considered flawed because there is no attempt at 'equating' in the Angoff sense.

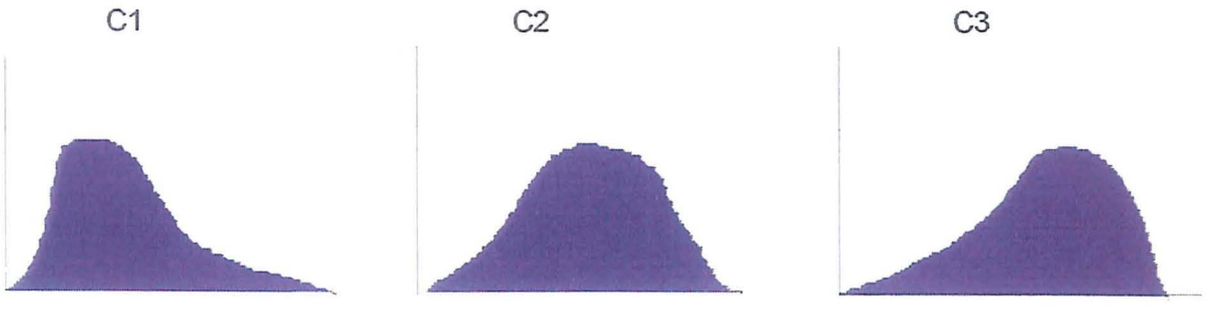
Component-Level Grading

For many years by choice, but most recently by dictat, public examination grade thresholds have been derived from the constituent components. Grading, or drawing the line, is done on component distributions, and it is this, currently judgemental, process which gives component marks their value by defining mark ranges for the various grades. There is no attempt to reconcile component distributions within linear schemes

of assessment so that like would be added to something more like. Aggregation can give rise to a number of potential paradoxes because the values ascribed to the marks that are added are different.

For example, suppose we have a three component examination, each of whose mark distributions exhibit different and often seen characteristics. Component 1 (C1) is positively skewed, component 2 has a reasonably normal distribution and component 3 is heavily negatively skewed. These mark frequency distributions are sketched below.

Figure B.1: Three Types of Mark Distributions



Each distribution can, in part, be described by the relative positions of the measures of central tendency, in particular:

for a positive skew:	mean > median
for a normal distribution:	mean = median
for a negative skew:	mean < median

and these have consequences for grading. The most obvious is that if, simplistically, one relates performance to mean score, candidates will actually do less well in terms of grade for negatively skewed distributions. In other words, given identical grade boundaries, the component grade distribution for a given group of candidates may vary markedly even when the performance mean is the same. It is, therefore, important to take into account the distributional characteristics of the raw marks in the grading process it is these on which the outcome of the aggregation process critically depends. Conversely, if syllabus grading is strictly norm-referenced, grade thresholds will be determined by the shape of the mark distribution even though the performance of the candidates, as exemplified by mean score, might suggest other factors may need to be considered.

The problems are exacerbated by the aggregation process. Since numerically, the grade boundaries from different distributions will be dissimilar even when found from norm-referencing, in combination the range of component grade profiles which produce the same syllabus grade can be vast. Two examples will serve to illustrate (details of which can be found in the annex to this appendix).

In the first, component grade profiles of B, A and A for the three components and D, C and B both give a syllabus grade of B. Note that these types of profiles could both result from well correlated papers. Whilst these may not be particularly close in mark terms they are both, by definition, of B standard yet they are also both the result of very different performances. It could be argued that on average B, A and A should produce at least an overall B grade, whereas D, C and B should average out at C, a one grade discrepancy. Of course, had the grade boundaries for each component been the same, such discrepancies could not have been found.

The second example shows grade profiles of A, A and D as against D, E and C both giving a syllabus grade of C. Again averaging the component grades would give expected syllabus grades of B and D respectively, in this case a two grade difference.

Such infelicities may be thought to be the result of data degradation i.e. reducing a raw mark scale to a single point, but if grades can be thought of as defining equal ranges on an interval scale, then the same anomaly would be found whatever the extent of that range (see later).

Both examples show the problems of aggregating raw marks without reference to the component distributions because the meaning, in terms of grade achievement, of the raw marks differ. A mark of 60 added to 80 will always produce 140 marks whether they represent performances of A and A, A and C or B and A. They are presumed to have an equal, undefined value and are aggregated on this basis. In fact raw scores are uncalibrated (i.e. have not been ascribed a 'value') and of little practical purpose on their own.

The setting of component grade boundaries is equivalent to ascribing to component marks a value, as well as defining syllabus grade boundaries. Thus, in the example quoted, 60 + 80 is not the same as 80 + 60 because the value of the marks, in terms of

grade, scored on one component is not the same as the value of the marks scored on another. The implications of this are profound, not least in defining where the standard of an examination must lie.

Conversions and UMS Scores

Because grade boundaries vary from session to session and module to module, candidates will have raw scores which have different meanings, especially in relation to the boundary values. The previous section attempted to highlight the difficulties of aggregating a three component examination when each component had different grade boundaries. Relative to these boundaries the raw marks took on, it is argued, different values. The problem with modular schemes when potentially aggregating six modules from four sessions and twenty two modules is considerably more complex and resolution is not easy. Individual syllabus boundaries are not the answer not only for practical reasons but for the fact that the same raw mark score will produce considerable anomalies in the grades awarded. For example, a candidate taking a resit, with a better result relative to the grade boundary, may find him/herself actually receiving a worse grade as a result of the aggregation process.

There are a number of techniques which may be used to overcome the problem of aggregating marks which are not to a common scale (as defined by the grade points). The setting of grades is arbitrary in the sense that, provided they are ordered, they can be set anywhere within the raw mark range and thus transformations of raw scores will not change significant parameters, especially the rank orders within components. (N.B. this is not true of a strictly criterion referenced mark scheme with pre-defined grade boundaries which, if it is to be robust, requires considerable pre-testing and adjustment and without this it is, in general, not to be encouraged). However, many transformations do change the variability of the scores and it is this variability which needs to be preserved for aggregation purposes. Artificially changing the dispersion of a component (which would be the effect of the usual statistical transformations) would give it a weight which it may not possess because variances would be affected.

It is acknowledged that grade boundary setting is a far from precise activity and that any imprecision will translate to the conversion system that is used. However, in the spirit of the assumption of error-free judgements which are implied by judgemental comparability, it is argued that in order to reduce anomalies which can be found when

'personal raw mark boundaries' are set, it is necessary to find a conversion scheme which is at least as fair as possible within the anomalies outlined in the previous appendix. Raw mark syllabus boundaries are as subject to error on a raw mark scale as on a converted scale.

There are a number of features which are desirable in any conversion. Assuming that conversions are to a common scale then that scale must reflect not only the grade awarded on the component, but how good that grade is. The dilemma with many conversions results from trying to decide how accurate they are. This is usually done with reference to the raw mark result which has already been shown to be a flawed indicator.

Thorndike and Hagen define three properties which units on a converted scale should have:

1. Uniform meaning from test to test, or in the case of GCE and GCSE, a uniform meaning from component to component or module to module within an examination (between subjects is a whole other ball game which will not be considered here)

2. Units of uniform size, i.e. equal interval scales

3. A true zero.

There are a number of possible conversions which conform to the above and could have been used, but following a comprehensive investigation by Thomson (1992), the uniform marks score (UMS) has been widely adopted. Accuracy here was determined by reference to grade profiles although these, by themselves, cannot be combined in a way which would reflect how good the grades were. (This is generally known as premature approximation when point scales are so much shorter than raw mark scales that relevant information is lost). Although UMS is usually applied in the modular context, it is argued that it is also a much fairer way of determining final scores for linear schemes of assessment than the current raw score method.

Again two examples (see annex for details) serve to illustrate the point. Two sets of profiles D, C and A or A, A and D give the same raw mark and therefore the same syllabus grade of C. Conversion to UMS would give the grades of C and B respectively.

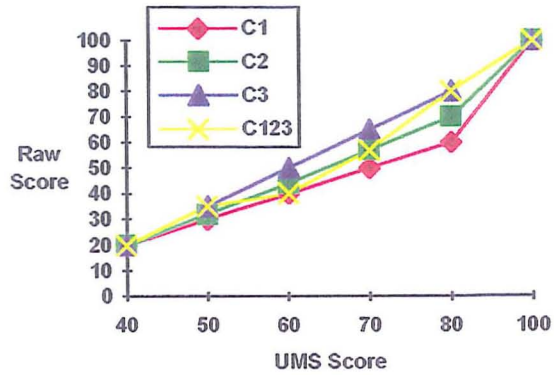
In fact it is possible to show in the same example that by adding a couple of marks to the C component mark, a higher syllabus grade is achieved from a lower raw mark.

The second example shows the component profiles of D, E and C or A, C and E giving the same raw mark and thus syllabus grade (C). The UMS scores suggest that a more accurate result would be syllabus grades of D and C. Intuitively, it would seem reasonable that equally weighted grades of D, E and C should average D and A, C and E should average C.

An interesting facet of UMS conversions is that strictly only raw scores above the nearest (lower) boundary are involved. Up to that bound, all raw scores convert to the same UMS score. Only raw scores above have to be scaled to fit the UMS band in order to show how good the grade is. Conversions are, therefore, only piecewise linear i.e. linear within a given grade. If all grade mark bandwidths are the same then the linearity will cross boundaries. This is illustrated in figure 2 where C1, C2 and C3, which were designed to have even grade bandwidths are shown to be linear, except for grade A, whereas C123, which is a composite with boundaries of 80 (A), 57, 40, 35 and 20 (E), is only linear between grade boundaries.

There are two reasons which make conversion of A grade raw marks shown here different. One is that the raw mark A range for all the examples is different from the other grade bandwidths and the other is that there are 20 UMS marks in the A range. Had both raw and UMS bands increased in the same proportion then linearity would have been continuous.

Figure B.2: UMS Conversions

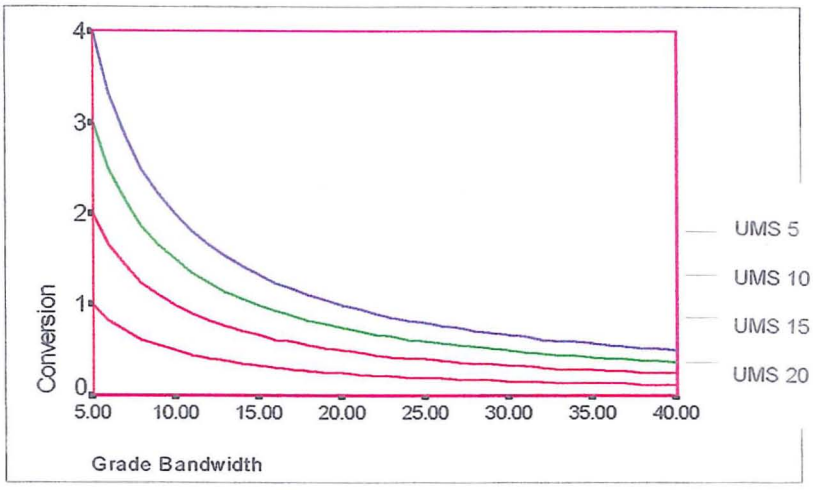


In addition to this piecewise linearity there is the exchange rate of raw to UMS score. Whilst, within a grade, all raw marks bear a constant relationship to each other, i.e.

$$UMS = K * RAW, \quad \text{where } K \text{ is constant within a grade,}$$

the conversion factor K varies in a distinctly non-linear fashion for different UMS ranges and raw mark grade ranges. This is illustrated in figure B.3 where graphs of K for different grade bandwidths are shown for four different UMS ranges.

Figure B.3: UMS Exchange Rates



The exponential decay in the conversion factor K has clear implications for varying grade bandwidths, and this is especially true of the A grade range where, for many subjects, the grade bandwidth is relatively large. (It is worth noting that the grade bandwidths B/C to D/E are arithmetically determined to be within one or two marks and it is rare to find an A/B range to be that different). The 'value' of a raw mark in the A grade range may not worth as much as a mark in a narrower grade band and there is a clear indication of diminishing returns as bandwidths increase. In this case the compensation from an A grade performance on one component will not be as great as expected from the raw mark (especially when compared with raw mark compensations where all marks are equally weighted). This is a particular problem, especially in the light of the desirable properties defined by Thorndike and Hagen.

Since in some subjects not all the mark range is used, possibly because of poor question setting or bad mark schemes or even tradition, the loss of compensatory power can be quite critical. Equally, in some subjects where all the raw mark range is used and candidates regularly score full marks, to over-compensate for an A scored on one component can be just as unfair.

Returning to the example of the annex, if 156 UMS had been scored from two of the components, 84 UMS would be required from the third in order to obtain an overall A. If this was to be gained from component 1 then 68 raw marks would be needed, if from component 3 then 84 raw marks would be needed. The important point to note is that twice as many raw marks above the A boundary would be needed on component 1 in order to gain the compensation required.

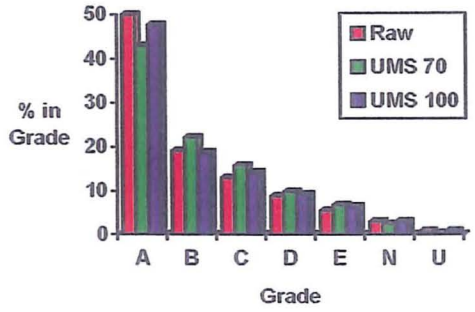
Effects of Grade Band Variations

There are then a number of issues which must be considered when conversions are suggested. UMS scores are generally used in modular A levels where module examinations may be taken on a number of different occasions. They are also used at GCSE when candidates may cross tiers for different parts of an examination, or again when a number of examining opportunities are available. They are also becoming the method of choice for aggregating marks when a number of different options are available e.g. History A level. The important point is that post-conversion, UMS boundaries (or indicator 1) are invariant and independent of option, tier or occasion.

However, the effects of choosing a UMS scale which is not suitable for a given syllabus can be profound, although once chosen it should not be subjected to year on year variation, although the conversion factor (K) from raw to UMS scores will change as raw score boundaries vary. What follows is an investigation into the changes in grade distribution that would occur in a cohort of candidates who have taken six modules in an A level examination under different conversion regimes. The raw scores for the candidates are unchanged, and are in fact well correlated, but in order to analyse conversion issues, other parameters have been artificially altered. Originally the raw score for each equally weighted module was out of 60, a score which several candidates achieved on some of the modules, though none on all of them.

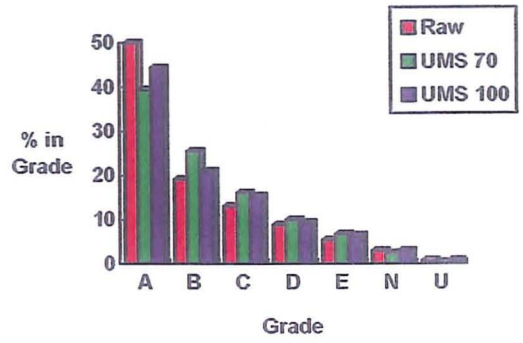
The data is a sub-set of 1801 candidates who have all taken the same combination of modules in the MEI Structured Mathematics examination. The module grade boundaries are initially those actually determined at the award. The UMS scale used for this syllabus is non-standard in that 60 is the A boundary, 50 the B and so on, with a maximum mark of 70. The more usual scale is to have 80 as the A bound, 70 as B etc. with a maximum of 100 marks. It has been long argued that because there is full use of the whole raw mark range, to double the UMS conversion from 10 to 20 for A will over-compensate candidates. This is certainly borne out by the results shown below where the number of A grades is increased by 5% when the UMS range for A is doubled.

Figure B.4: Effect of Change in UMS Scale



Suppose the marks are kept the same, but the maximum raw mark for each module is taken as 70 instead of 60. This simulates a poorly discriminating set of papers, or a subject where the whole of the mark range is not employed. The results for this are shown below.

Figure B.5: Effect of Change in Maximum Raw Mark

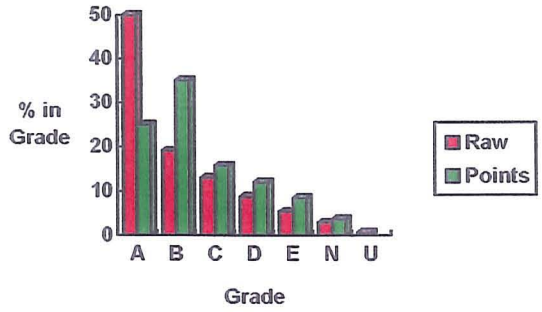


Comparison with the same raw results shows that the effect of not using the whole mark range is marked with about a 3% reduction in the candidates obtaining an overall A.

A more dramatic difference is apparent when points are awarded for each grade (10 for an A, 8 for a B and so on), when any compensatory effect from a longer raw mark A range is completely lost. The results are shown below, again contrasted with the same raw results.

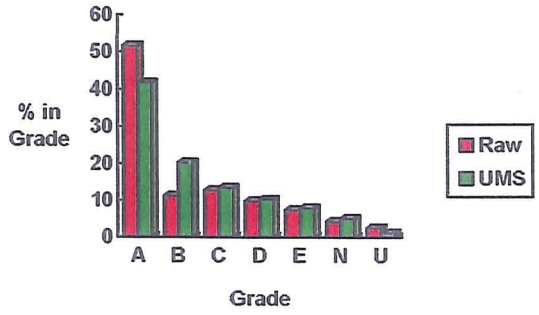
The percentage of A grades using this calculation is halved, with a concomitant increase mainly in the B percentage. Such a drastic loss in compensatory power is considered undesirable especially when, as in this case, the raw mark A range is approximately 3 times that of other grade bandwidths. Whereas it is conceivable that, for this syllabus, raw mark scores tend to over-compensate at the top end, to base awarding on equal UMS grade bandwidths is not recommended.

Figure B.6: Effect of Grade Points Conversion



These analyses have not been systematic as, in a properly conducted syllabus (i.e. one that takes into account raw to UMS conversion characteristics), major differences in grade boundaries between modules would not be expected, nor would significantly different distributions. However, it is possible to simulate the effect of such differences. Using the same set of raw scores, the grade boundaries from each module have been norm referenced i.e. each boundary is derived from the distributional characteristics of the module in question according to the rule 50% obtain A, 60% B, 70% C, 80% D and 90% gain E or better. This rule produces a very different set of boundaries from those derived from judgement and widely differing values for K within and between modules. The results of this procedure are shown graphically below:

Figure B.7: The Effect of Norm-Referenced Boundary Values Between Modules

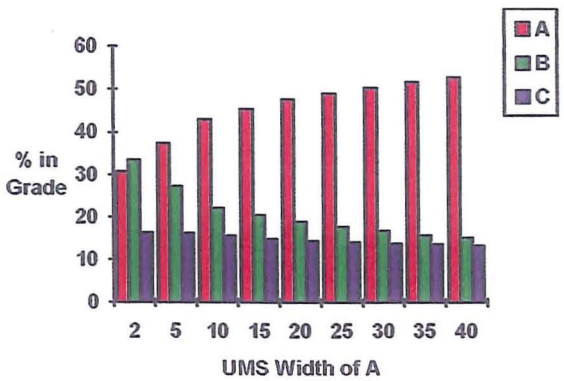


It is somewhat perverse, but it is in the nature of modular schemes that a fairly volatile module population is to be expected and norm-referencing at module level would not be recommended because the quality of candidates in any percentile range will not be constant between sessions. It can sensibly be used in this case only because the candidates have been deliberately chosen from those who have sat common papers. They are thus not only a subset of the syllabus candidates but also a subset of these particular module candidates, each of whose total populations differ from each other. Whilst no reliance can be placed on the fairness of the raw score derived distribution, there are clearly differences between what would be current practice for a linear scheme and the UMS converted distribution.

The Width of 'A'

So which is right? As in most awarding situations there is probably no right or wrong answer, but much of the discussion revolves around the conversion of A. Although the dataset used for this investigation may not be typical of each and every modular scheme, there should be enough commonality to enable the identification of trends with some accuracy. Although the A boundary for all modules is not the same, it does not vary widely and is typically of the order of 18 raw marks, which is about three times as great as every other module raw mark grade bandwidth. The straight raw mark aggregation gives 50% of the cohort an A aggregate for the syllabus. If we assume that all other grade conversions are to 10 UMS (which importantly is always greater than the raw mark range or premature approximations will occur), the effect of varying the width of the UMS conversion for A from 2 to 40 on the syllabus grade distribution is shown below:

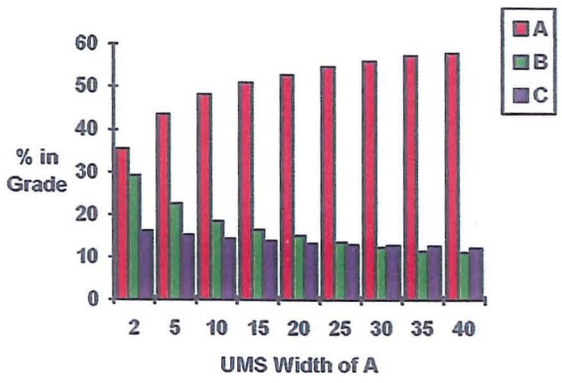
Figure B.8: Effect of Varying UMS Score for A



Obviously as the number of grade As increase, the number obtaining B will fall and there is a slight, but perceptible knock-on effect in the numbers obtaining a C. At lower grades there is no significant difference in percentages. Even with a conversion of 25 UMS marks, there are still marginally fewer candidates obtaining an A than apparent from the raw score. This is because the longer than raw score UMS conversion at lower grades requires that the compensatory power of A be greater. There is a trade-off between the UMS range at A and those further down the grades. And this must always be so since the salient feature of UMS scores is that they are invariant and independent of examining session, whereas raw scores are not.

Figure B.9 illustrates the point well. In this case the UMS ranges for all grades, except A, is halved to 5 marks. This is generally closer to the raw mark ranges, though generally a point or two smaller. Again varying the width of A a similar pattern emerges, but in this case the 50% percentage at A is reached at a conversion of 15 rather than the 30 UMS of figure B.8.

Figure B.9: Effect of Halving UMS Ranges



Whilst there is no implicit reason why 50% should be in any sense more accurate than any other figure, it does give some benchmark against which to gauge other effects. Although actual percentages vary with different UMS ranges, the changes in syllabus distribution follow the same trend as the A UMS range changes.

Discussion

Because of the varying boundary value for each grade, there can be no hard and fast rule for conversions. For example, based on the previous analysis, it is desirable that the UMS scale should be of the same order as the raw mark grade bandwidth i.e. conversions should be approximately in the ratio of 1:1. The former is invariant, the latter, over the period of the examination, not. Therefore such a policy cannot be executed with any rigour. However, this is not true of linear examinations where it may be sensible to use different UMS scales in different years if component distributions warrant it. Even in the linear case there is likely to be a compromise in order to choose a sensible scale for all components.

There are therefore two separate, and possibly conflicting, requirements of any conversion:

(i) equating unlike distributions and boundaries in order to adjust raw scores
(ii) equating distributions and boundaries in order to aggregate across time.

Assuming that UMS scores continue to be the method of attaining these goals then the scales used are likely to be a compromise between raw scales on the different

modules/components since all the indications suggest that matching raw to UMS mark bandwidths for each grade is likely to reduce the occurrence of anomalies. This would also suggest that UMS scales should be syllabus dependent and not common across all subjects. The characteristics of raw score distributions vary markedly and to use only one UMS scale irrespective of the raw score characteristics is unwise, although once chosen for a subject (syllabus) it cannot be changed because of the temporal aspect of modular schemes. The injudicious choice of UMS scale can clearly have a deleterious effect on the syllabus grade distribution, and it would be disingenuous to believe that module boundaries were not set with the consequences at syllabus level a consideration. As awarders become more aware of the consequences of their decisions in this two stage aggregation process, mark schemes and boundaries are being constrained by their potential effects.

However, an additional point which is important in the context of modular examinations, is the independence of each of the modules. Since grading takes place at module level, and strictly session-on-session comparability is between modules, not syllabuses, it is within the modules that grading standards lie. There is an opposite opinion which would stress the importance of the syllabus standard, but, providing the conversion to a time-invariant scale is sensible at module level, and UMS marks retain the qualitative power of the raw marks in the graded scale, there should not be too many anomalies thrown up by the aggregation of those modules.

If little else is gleaned from the foregoing analysis, it is clear that final grade distributions are extremely sensitive to the choice of scale. Expected syllabus grade distributions can be predicted from a number of performance indicators: last year's distribution, entry by centre and gender, benchmark centres, forecast grades, predictions from GCSE results. It is entirely possible, given a judicious choice of conversion factor, to produce a final grade distribution which is in line with performance indicators. The corollary to such an assertion would be that the standard of an examination was embodied in this choice.

The attainment of a syllabus grade is a two stage process, and this appendix addresses only the second stage, that of aggregation. Arguably the more important step is that of 'drawing the line', that is the determination of the various grade boundaries at module level.

Aggregation is mechanistic although it can undoubtedly have a profound effect on the final grades awarded to some individuals. What is of fundamental importance is that aggregation should be a process where like is aggregated with like. As has been shown, raw scores, by themselves, are, or can be, misleading because they have not been (non-statistically) 'standardised'. In awarding terms this means that they have not been calibrated with reference to grades. Whilst it has been seen as essential in modular awarding to relate raw scores to a common system, it is not now usually considered to be a cardinal element in the aggregation of linear schemes. There has always been a school of thought that the aggregation process actually disguises true performance, especially when the different components of an examination address very different assessment objectives. However, although tradition may be gradually overcome by the reporting of separate module results, there is little doubt that, to most observers and interested parties, attainment is still only considered meaningful at syllabus level. There is therefore also the feeling that this is where the standard of an examination lies.

It is argued that such an opinion is unsustainable in the light of the reconciliation process which is part of the aggregation of any multi-dimensional examination. Candidates with ostensibly the same performance at syllabus level, may have demonstrated very different trait attainment. Standards must lie at the level of value ascription, and that is with the setting of boundary values at component/module level. Once the boundaries have been drawn, each mark has been contextualised i.e. given a value. By its position within a grade band, a mark describes not only that the performance merits a given grade but also informs how good that grade is. No such value can be seen as attached to an aggregate mark because it is an artefact of the aggregation process which is itself uncalibrated.

The process of determining grade boundary values at component/module level is a complex one and is usually a compromise between judgement and statistics, with the former, under the current SCAA Codes of Practice, predominating. At this stage it is theoretically possible to take into account the demand of the question papers, the vagaries of the mark scheme and to weigh these against the evidence of attainment from the written papers. Although this is an imprecise and error-prone process, it appears to be acceptable to users of the qualifications. Holistic grading, which is generally against the Code of Practice, would entail weighing up very different demands and attainments from the various strands of the examination, each with its own distributional characteristics, and determining a boundary mark. However, this would

undoubtedly assume raw mark aggregation across components and, as has been shown, this may be a flawed process.

It follows that if the standard of an examination lies at component/module level, then it does, in a sense, define a syllabus standard because (provided regression is not an issue) a candidate who has obtained a given grade boundary mark on each component must also obtain a syllabus boundary mark at the same grade with respect to any reasonable conversion. The addition (or subtraction) of a component also on the boundary would not change the syllabus standard - although it would probably affect the demand of the examination by narrowing the domain of assessment, and possibly behaviour.

Key to the discussion of where standards lie is the issue of norm referencing. Whilst it may be possible to argue that, in the final analysis, syllabus grades are influenced by statistical comparisons with previous years, as they must be to retain year-on-year comparability, it is the derivation of the syllabus mark distribution which is of fundamental importance. This distribution is critically affected by the choice, not only of units (which may be raw scores or other types of scale such as UMS) by which components/modules are aggregated, but crucially by the position of the grade thresholds on the raw mark scale. It is this latter factor which is the cornerstone of the awarding process, and sets the standard on which all other processes rely. A consequence of this may be that if module standards are deemed correct then there should be the flexibility to change the UMS syllabus boundary if year-on-year syllabus standards are not to be broached. However, it would be counter to the rationale behind UMS scores and produce the same problems as with raw scores i.e. the relationship of each module score to the syllabus total will vary and will depend upon other candidates who are also aggregating. This would not be fair.

ANNEX to APPENDIX B

Some Inconsistencies

Imagine an examination with three components C1, C2 and C3. Each of these components is equally weighted and carries a maximum mark of 100. The results of the examination lead to the establishment of three different, though evenly spread, sets of boundary marks given below.

	A	B	C	D	E
C1	60	50	40	30	20
C2	70	57	44	32	20
C3	80	65	50	35	20
Total	210	172	134	97	60

Assume too that each component grade can be converted to a point score with 10 for an A, 8 for a B, 6 for a C, 4 for a D and 2 for E.

Example 1 - Syllabus Grade B

	C1	C2	C3	C1	C2	C3	Total	Grade	Point Average
Profile 1	50	70	80	B	A	A	200	B	9.3
Profile 2	38	56	79	D	C	B	173	B	6.0

Example 2 - Syllabus Grade C

Profile 3	60	70	35	A	A	D	165	C	8.0
Profile 4	39	31	64	D	E	C	134	C	4.0

Suppose additionally that a standard UMS conversion is assumed i.e. 80+ for an A, 70-79 for a B, 60-69 for a C, 50-59 for D and 40-49 for an E with a maximum score of 100. The mappings for each component boundary from raw to UMS scores are thus:

C1 Raw	C2 Raw	C3 Raw ---->	UMS
60	70	80	80
50	57	65	70
40	44	50	60
30	32	35	50
20	20	20	40

Two further profiles are defined:

Profile 5	38	46	81	D	C	A	165	C	6.7
Profile 6	60	54	20	A	C	E	134	C	6.0

Comparing UMS conversions of profiles 3 and 5, the same raw score will lead to UMS scores of 210 and 200 respectively with corresponding grades of B and C, the same as the point average.

Comparing UMS conversions of profiles 4 and 6, the same raw score will lead to UMS scores of 177 and 208 respectively with corresponding grades of D and C respectively, again the same as the point average would indicate.

There are clearly a large number of similar examples which would serve to illustrate the same point. Equally clearly if correlations are high between components and mark distributions very similar, far fewer dichotomies of the type illustrated would exist.

However, such simple examples also serve as a reminder that aggregation of raw scores is flawed and that any discussion of the merits of different types of conversions should not rely entirely on comparisons with raw scores.

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

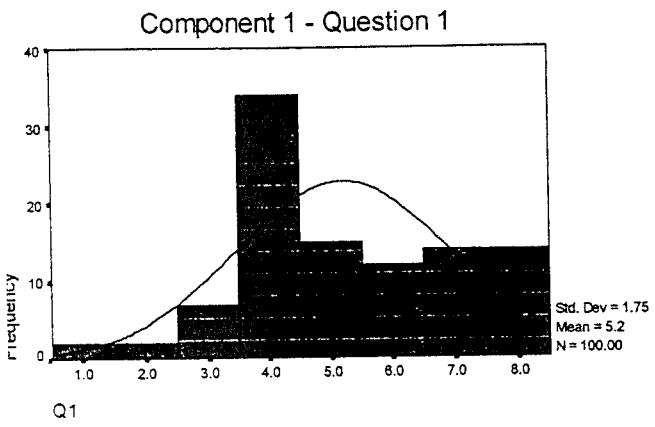
TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

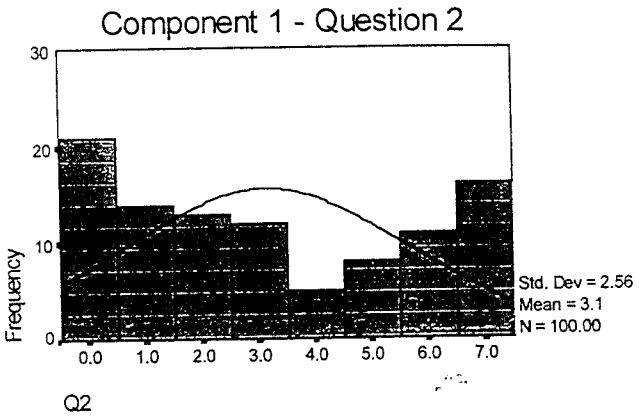
TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

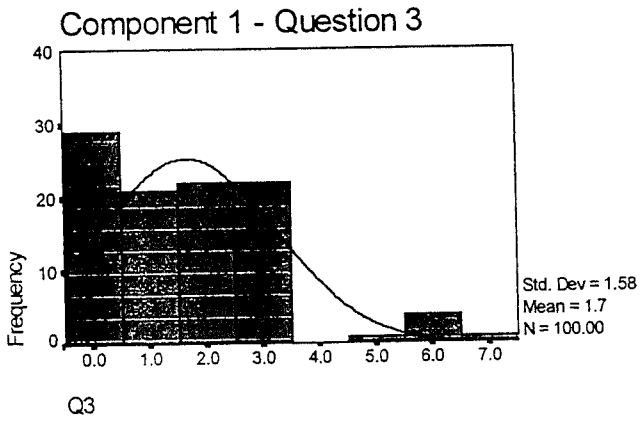
TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

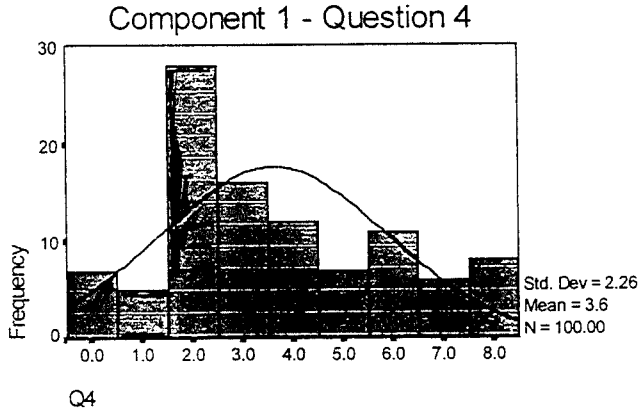
TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES

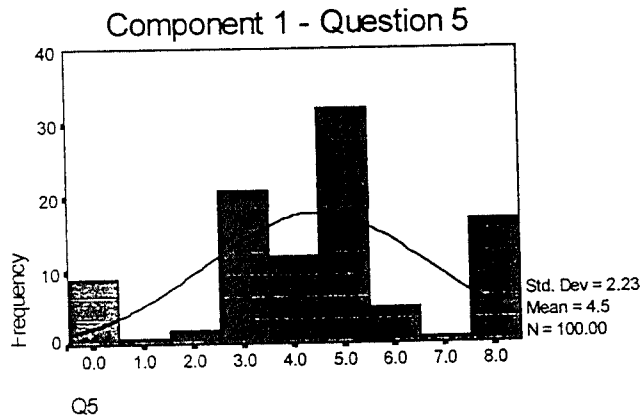
Question Level Data

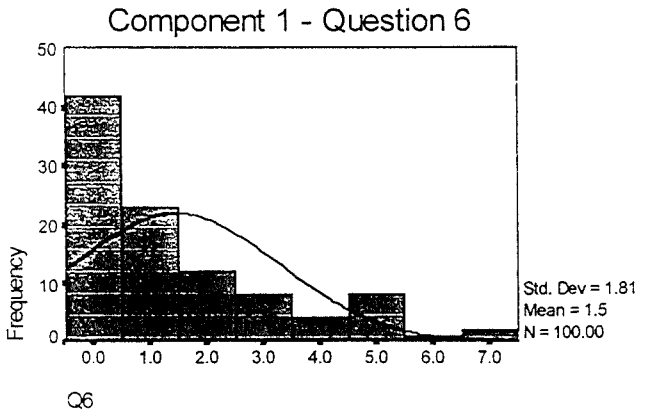


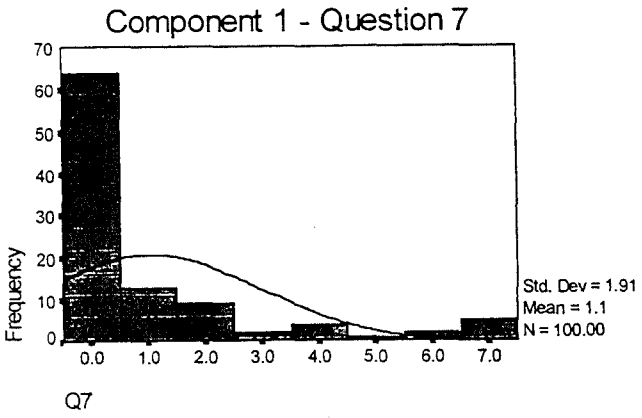


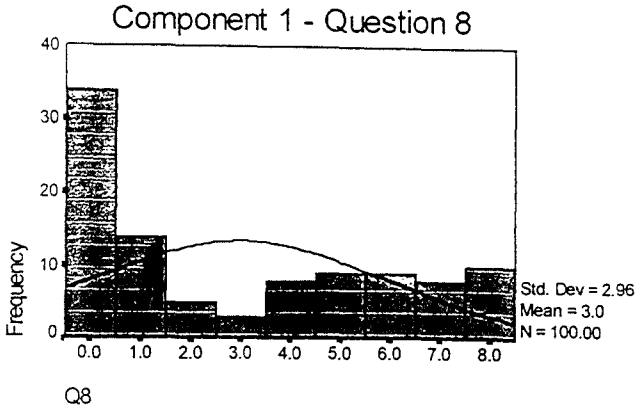


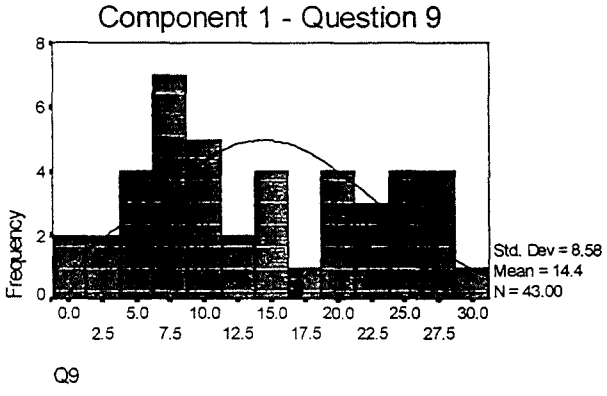


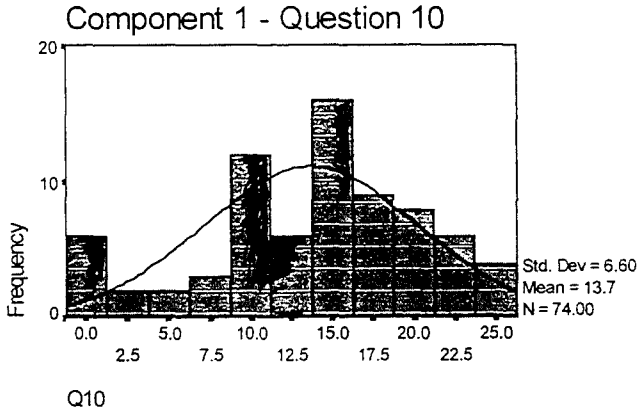


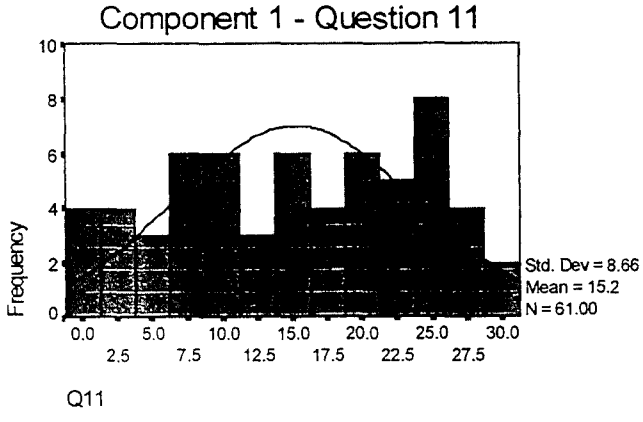


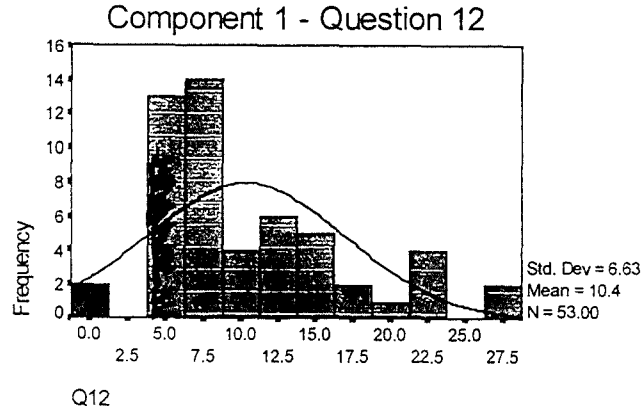

Component 1 - Question 7
Frequency
Std. Dev = 1.91
Mean = 1.1
N = 100.00
Q7

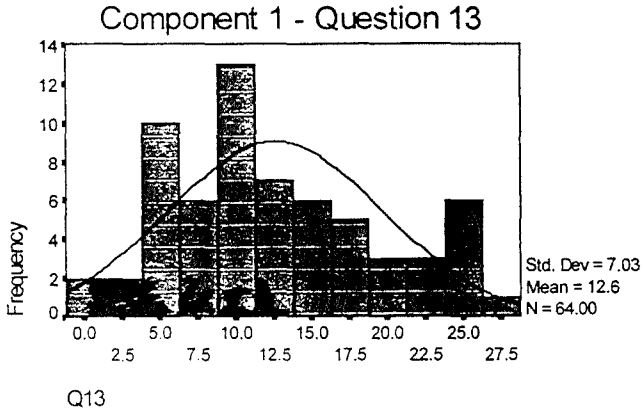

Component 1 - Question 8
Frequency
Std. Dev = 2.96
Mean = 3.0
N = 100.00
Q8

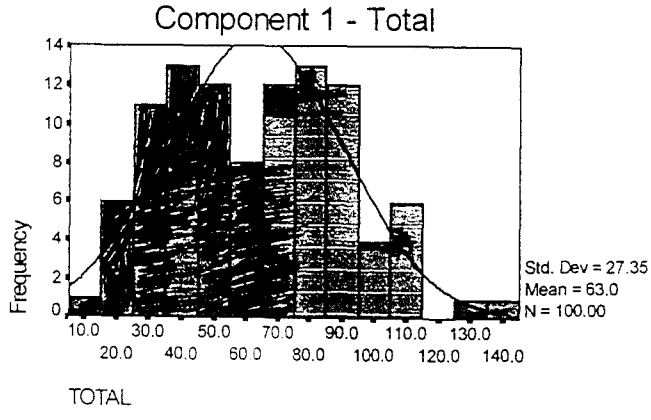

Component 1 - Question 9
Frequency
Std. Dev = 8.58
Mean = 14.4
N = 43.00
Q9

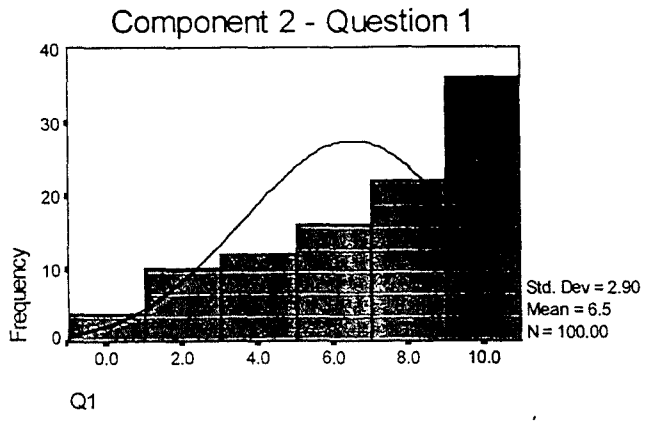

Component 1 - Question 10
Frequency
Std. Dev = 6.60
Mean = 13.7
N = 74.00
Q10

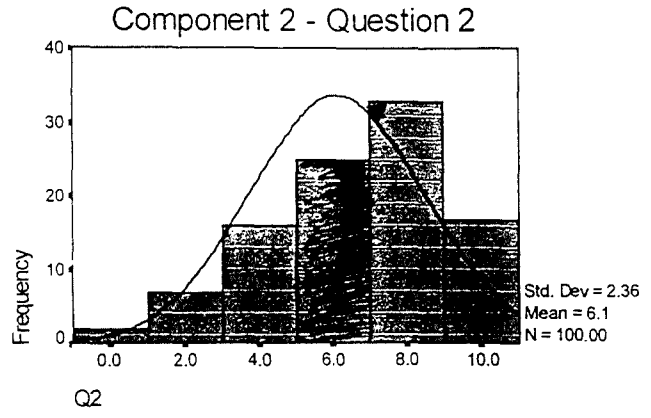

Component 1 - Question 11
Frequency
Std. Dev = 8.66
Mean = 15.2
N = 61.00
Q11

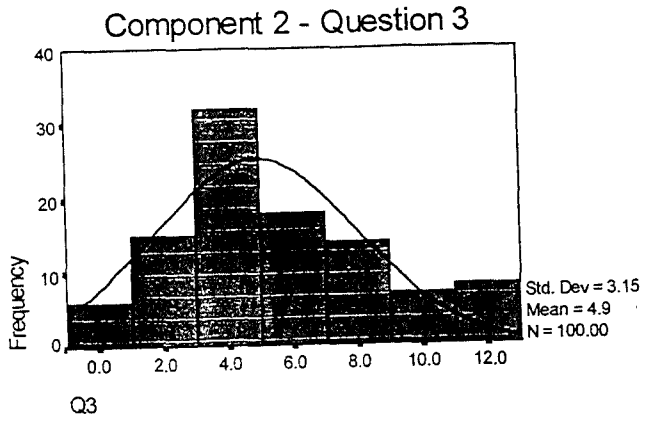

Component 1 - Question 12
Frequency
Std. Dev = 6.63
Mean = 10.4
N = 53.00
Q12

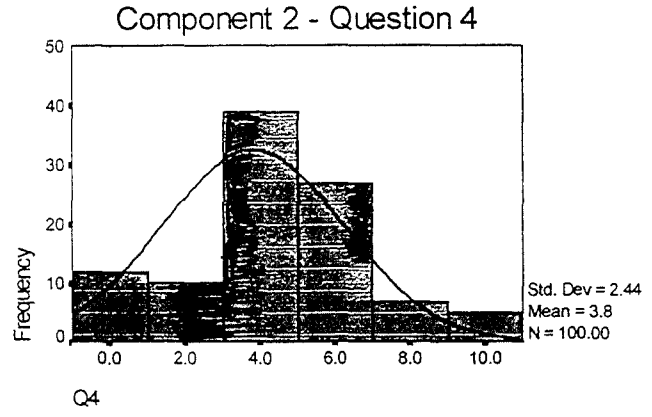

Component 1 - Question 13
Frequency
Std. Dev = 7.03
Mean = 12.6
N = 64.00
Q13

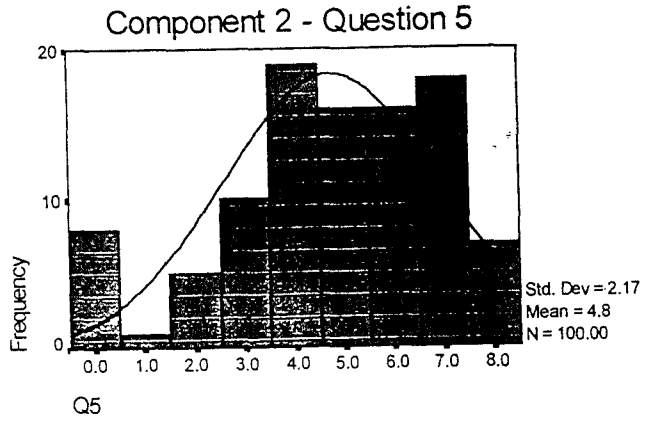

Component 1 - Total
Frequency
Std. Dev = 27.35
Mean = 63.0
N = 100.00
TOTAL

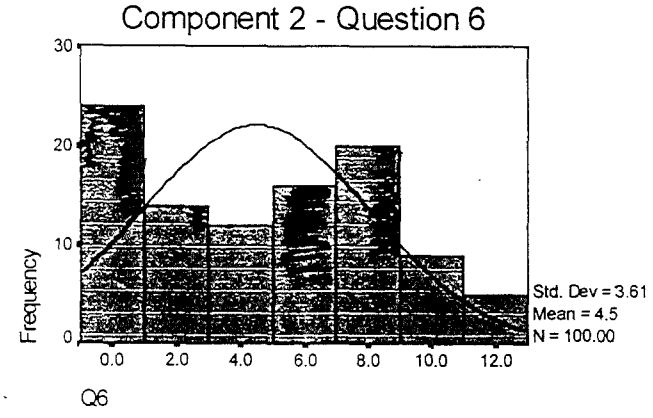

Component 2 - Question 1
Frequency
Std. Dev = 2.90
Mean = 6.5
N = 100.00
Q1

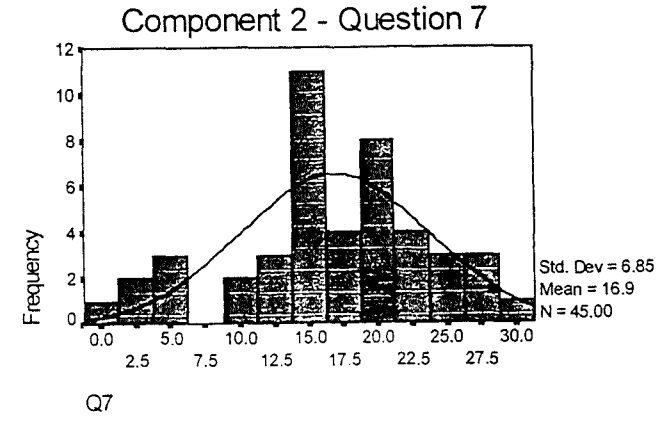

Component 2 - Question 2
Frequency
Std. Dev = 2.36
Mean = 6.1
N = 100.00
Q2

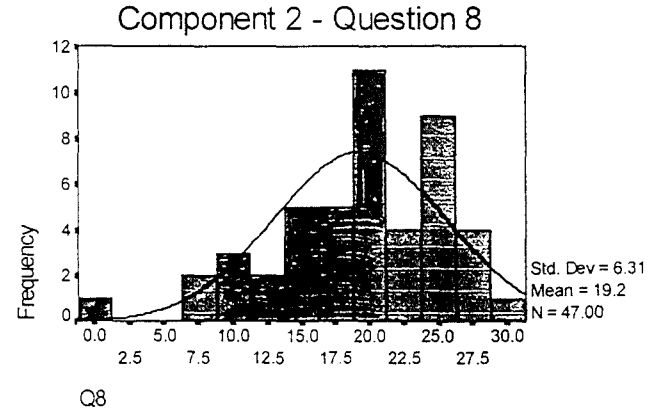

Component 2 - Question 3
Frequency
Std. Dev = 3.15
Mean = 4.9
N = 100.00
Q3

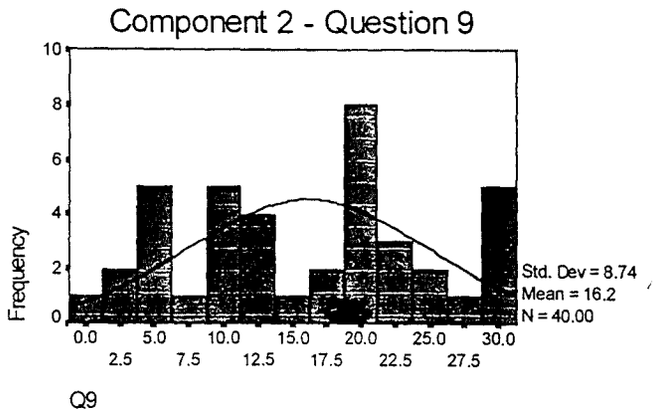

Component 2 - Question 4
Frequency
Std. Dev = 2.44
Mean = 3.8
N = 100.00
Q4

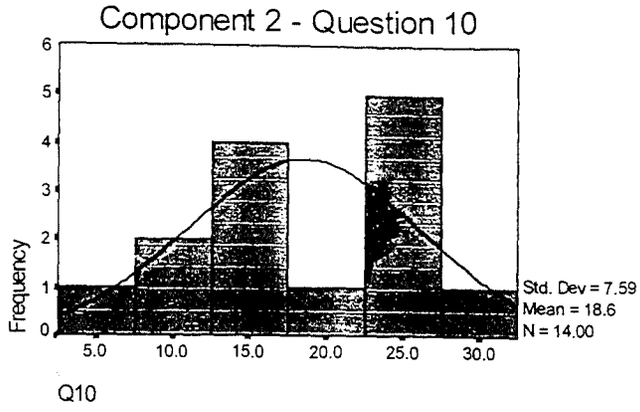

Component 2 - Question 5
Frequency
Std. Dev = 2.17
Mean = 4.8
N = 100.00
Q5

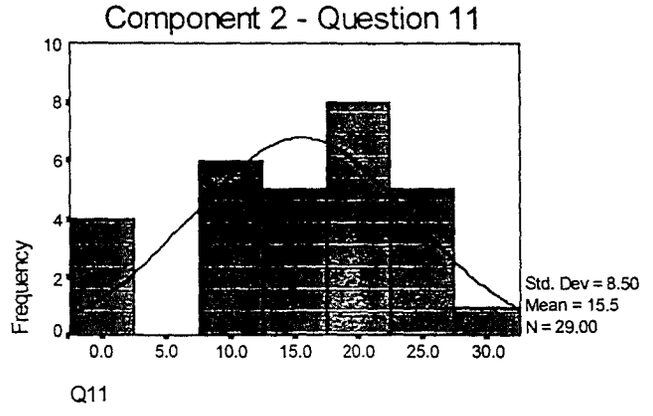

Component 2 - Question 6
Frequency
Std. Dev = 3.61
Mean = 4.5
N = 100.00
Q6

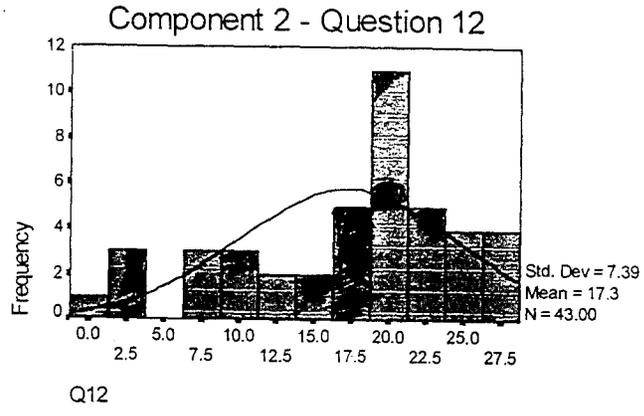

Component 2 - Question 7
Frequency
Std. Dev = 6.85
Mean = 16.9
N = 45.00
Q7

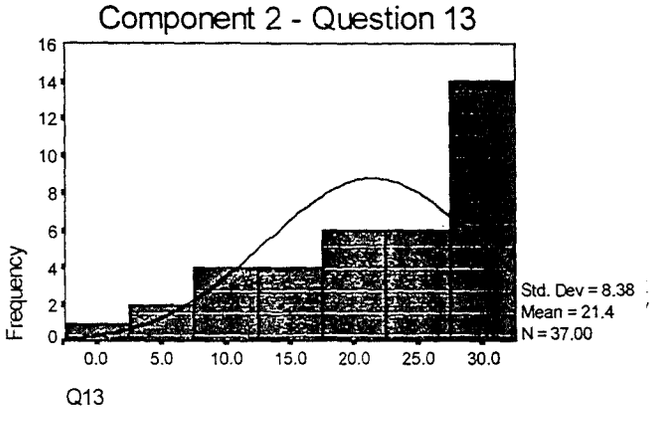

Component 2 - Question 8
Frequency
Std. Dev = 6.31
Mean = 19.2
N = 47.00
Q8

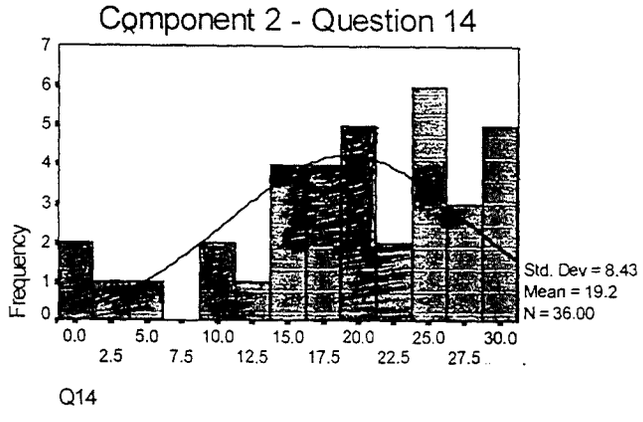

Component 2 - Question 9
Frequency
Std. Dev = 8.74
Mean = 16.2
N = 40.00
Q9

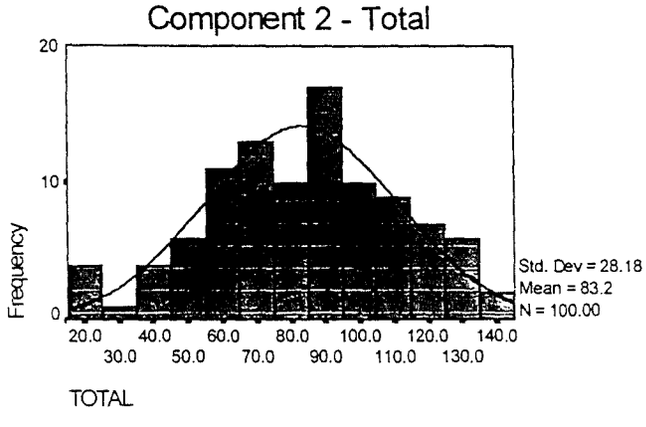

Component 2 - Question 10
Frequency
Std. Dev = 7.59
Mean = 18.6
N = 14.00
Q10

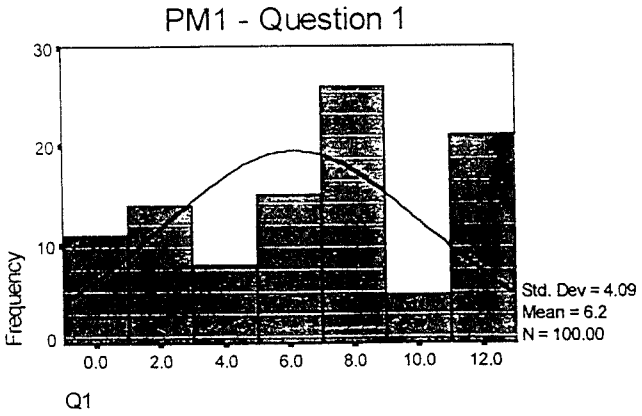

Component 2 - Question 11
Frequency
Std. Dev = 8.50
Mean = 15.5
N = 29.00
Q11

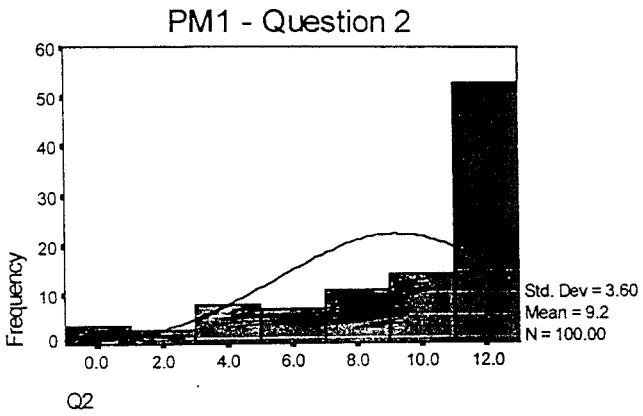

Component 2 - Question 12
Frequency
Std. Dev = 7.39
Mean = 17.3
N = 43.00
Q12

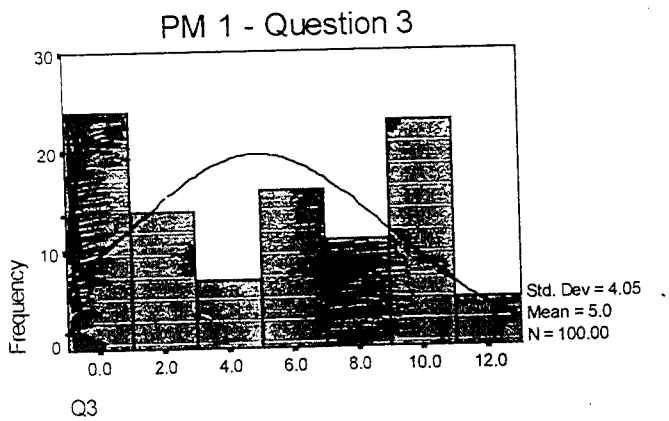

Component 2 - Question 13
Frequency
Std. Dev = 8.38
Mean = 21.4
N = 37.00
Q13

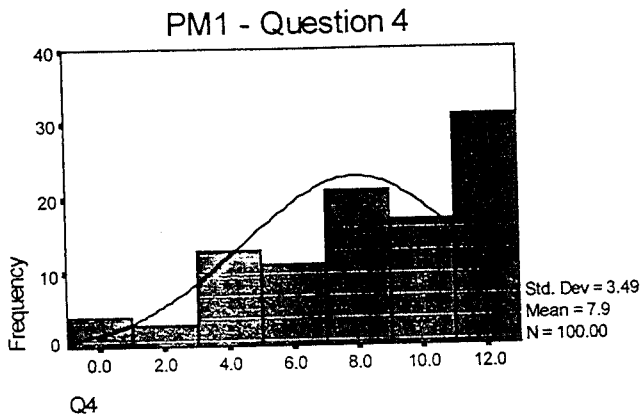

Component 2 - Question 14
Frequency
Std. Dev = 8.43
Mean = 19.2
N = 36.00
Q14

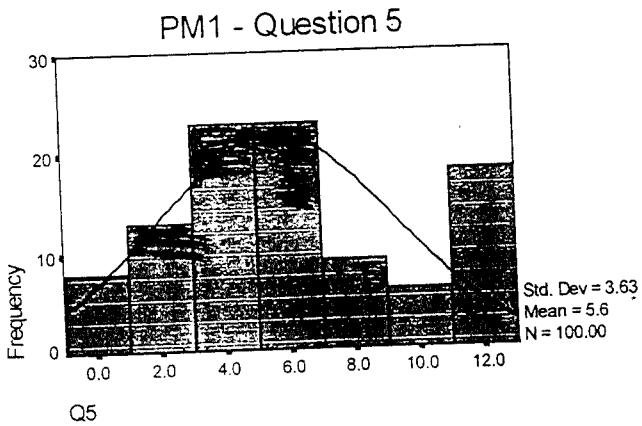

Component 2 - Total
Frequency
Std. Dev = 28.18
Mean = 83.2
N = 100.00
TOTAL

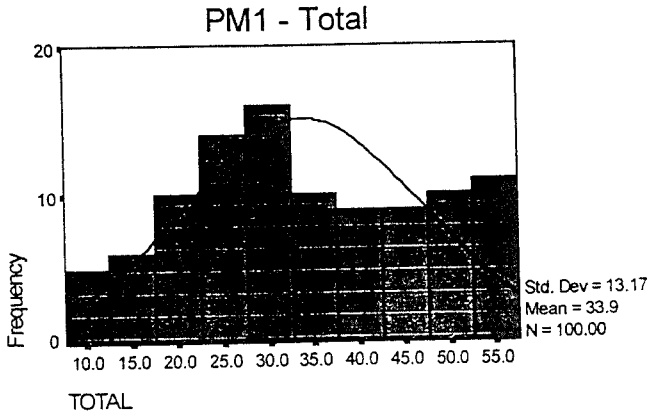

PM1 - Question 1
Frequency
Std. Dev = 4.09
Mean = 6.2
N = 100.00
Q1

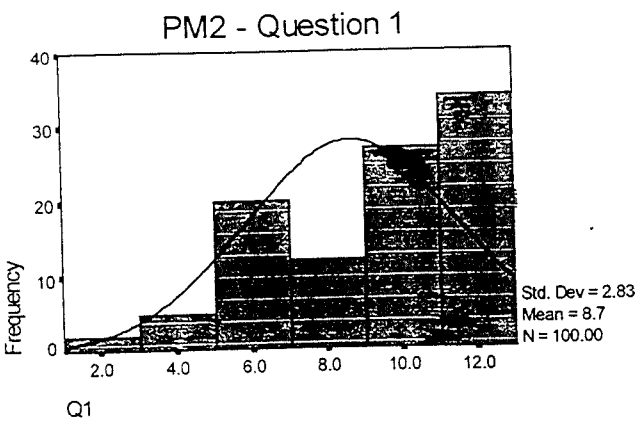

PM1 - Question 2
Frequency
Std. Dev = 3.60
Mean = 9.2
N = 100.00
Q2

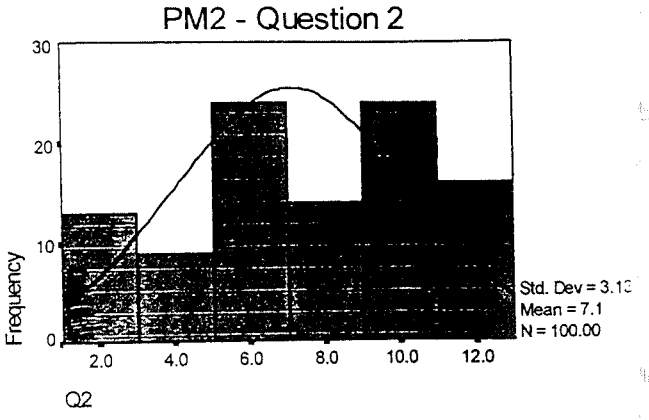

PM 1 - Question 3
Frequency
Std. Dev = 4.05
Mean = 5.0
N = 100.00
Q3

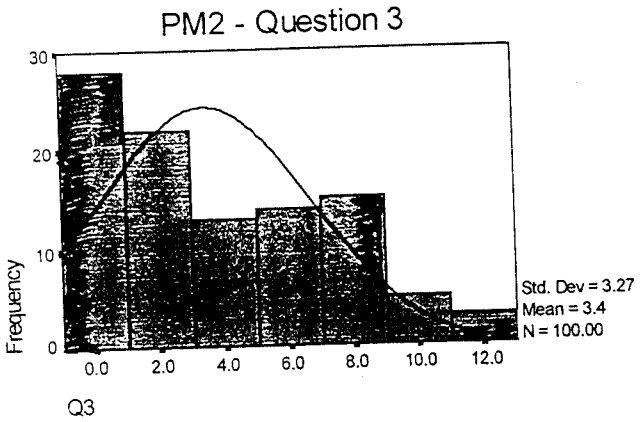

PM1 - Question 4
Frequency
Std. Dev = 3.49
Mean = 7.9
N = 100.00
Q4

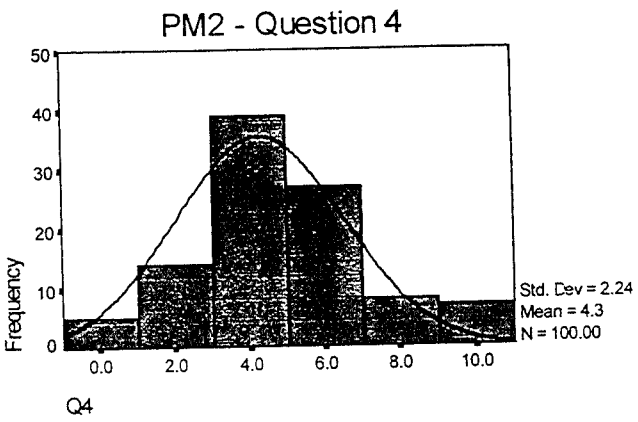

PM1 - Question 5
Frequency
Std. Dev = 3.63
Mean = 5.6
N = 100.00
Q5

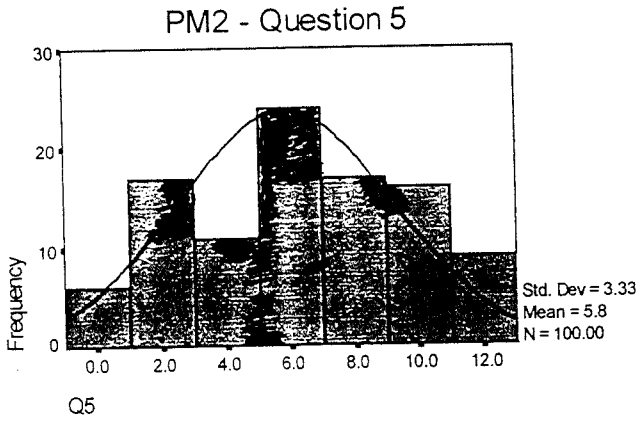

PM1 - Total
Frequency
Std. Dev = 13.17
Mean = 33.9
N = 100.00
TOTAL

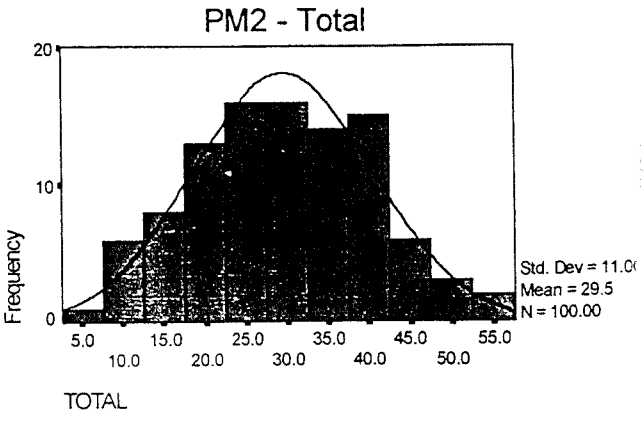

PM2 - Question 1
Frequency
Std. Dev = 2.83
Mean = 8.7
N = 100.00
Q1

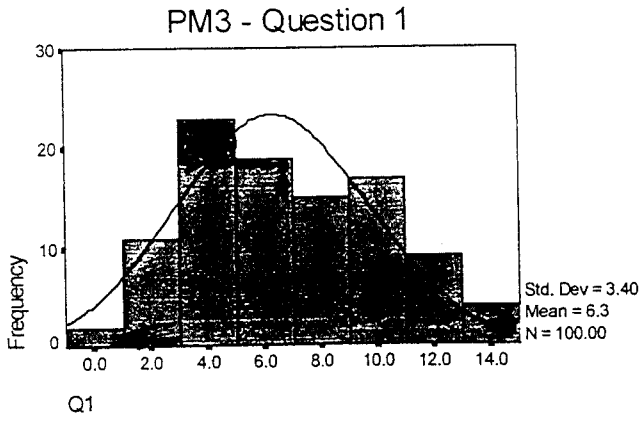

PM2 - Question 2
Frequency
Std. Dev = 3.1
Mean = 7.1
N = 100.00
Q2

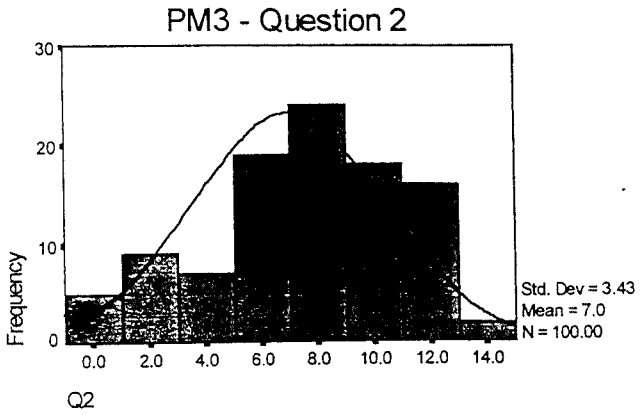

PM2 - Question 3
Frequency
Std. Dev = 3.27
Mean = 3.4
N = 100.00
Q3

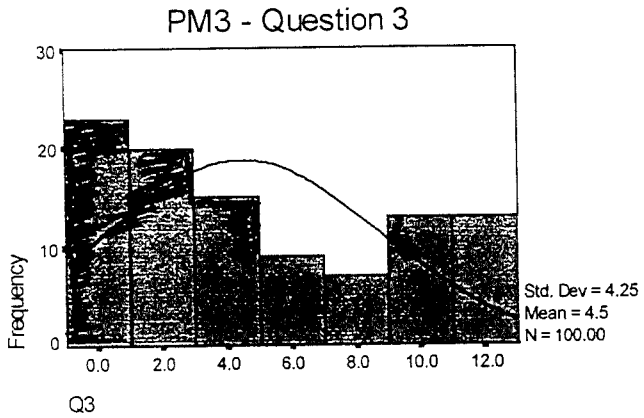

PM2 - Question 4
Frequency
Std. Dev = 2.24
Mean = 4.3
N = 100.00
Q4

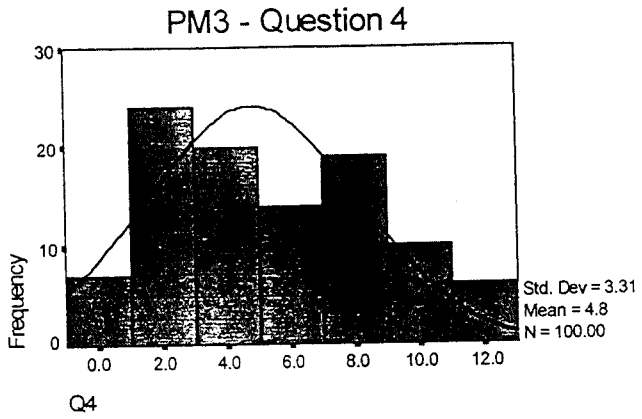

PM2 - Question 5
Frequency
Std. Dev = 3.33
Mean = 5.8
N = 100.00
Q5

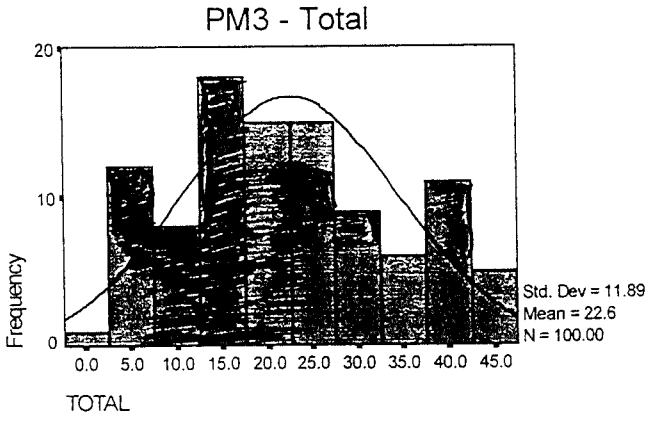

PM2 - Total
Frequency
Std. Dev = 11.0
Mean = 29.5
N = 100.00
TOTAL

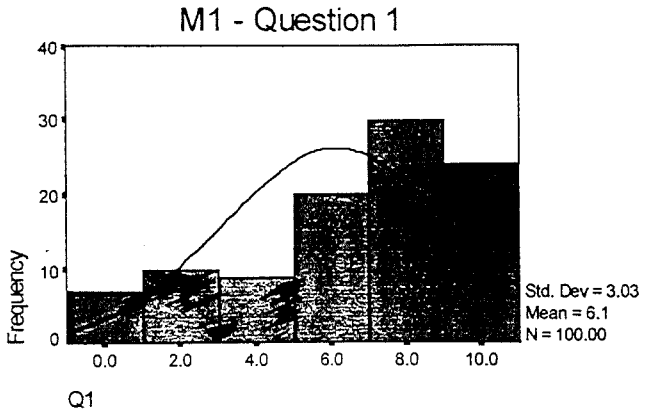

PM3 - Question 1
Frequency
30
20
10
0
0.0 2.0 4.0 6.0 8.0 10.0 12.0 14.0
Q1
Std. Dev = 3.40
Mean = 6.3
N = 100.00

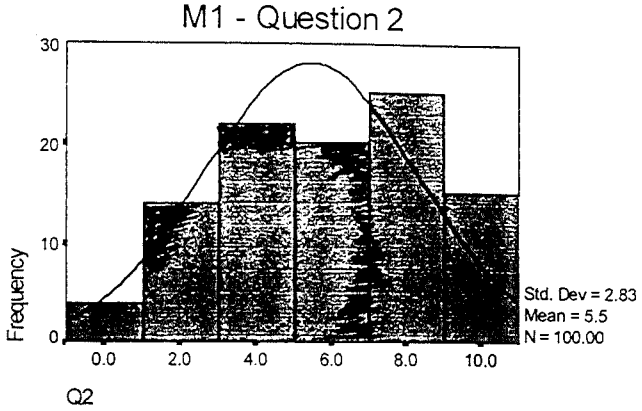

PM3 - Question 2
Frequency
30
20
10
0
0.0 2.0 4.0 6.0 8.0 10.0 12.0 14.0
Q2
Std. Dev = 3.43
Mean = 7.0
N = 100.00

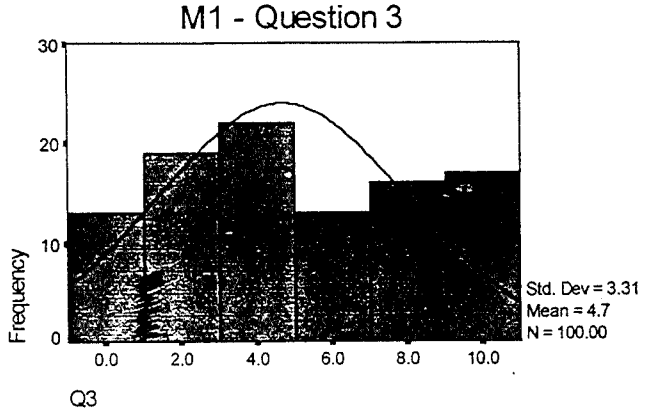

PM3 - Question 3
Frequency
30
20
10
0
0.0 2.0 4.0 6.0 8.0 10.0 12.0
Q3
Std. Dev = 4.25
Mean = 4.5
N = 100.00

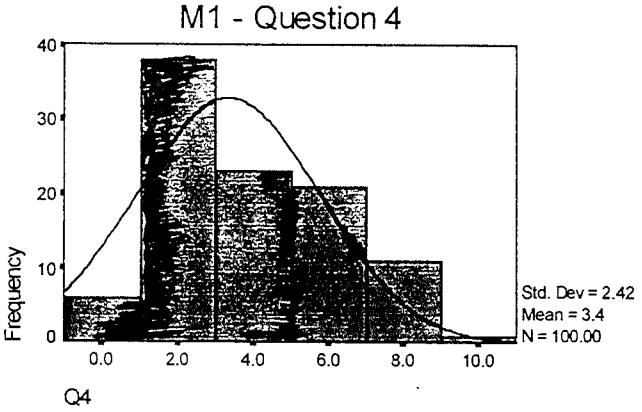

PM3 - Question 4
Frequency
30
20
10
0
0.0 2.0 4.0 6.0 8.0 10.0 12.0
Q4
Std. Dev = 3.31
Mean = 4.8
N = 100.00

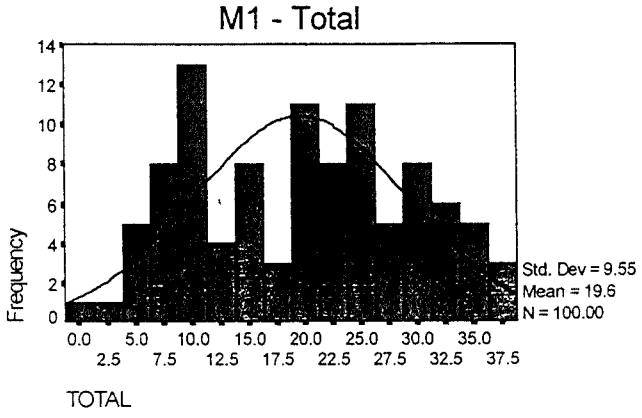

PM3 - Total
Frequency
20
10
0
0.0 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0 45.0
TOTAL
Std. Dev = 11.89
Mean = 22.6
N = 100.00

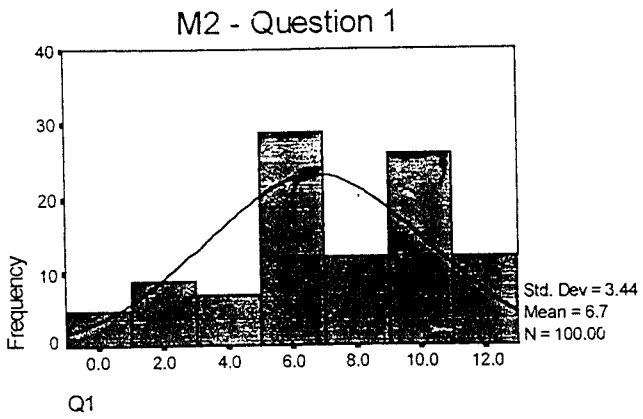

M1 - Question 1
Frequency
Std. Dev = 3.03
Mean = 6.1
N = 100.00
Q1

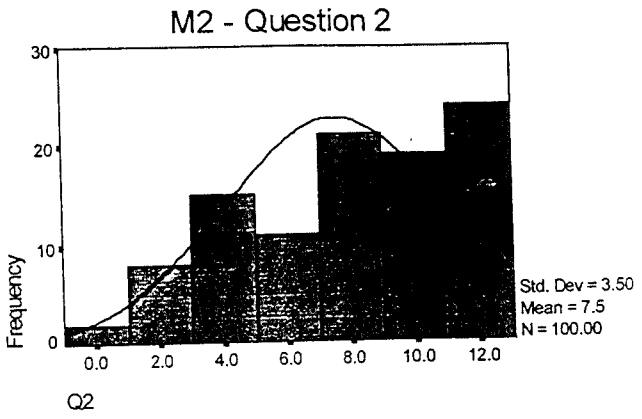

M1 - Question 2
Frequency
Std. Dev = 2.83
Mean = 5.5
N = 100.00
Q2

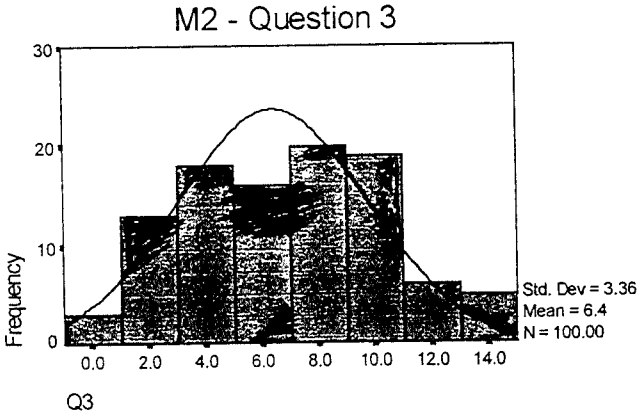

M1 - Question 3
Frequency
Std. Dev = 3.31
Mean = 4.7
N = 100.00
Q3

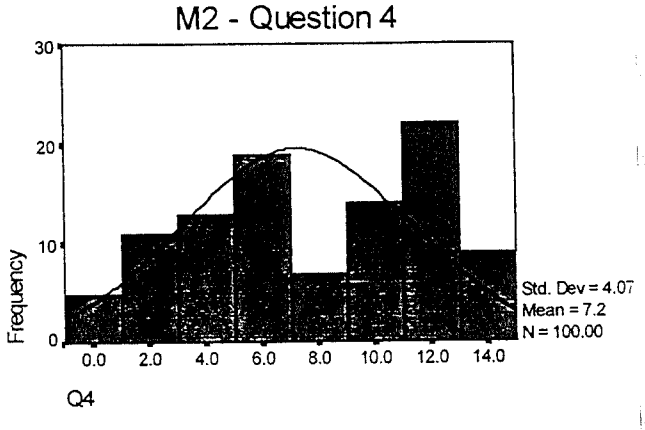

M1 - Question 4
Frequency
Std. Dev = 2.42
Mean = 3.4
N = 100.00
Q4

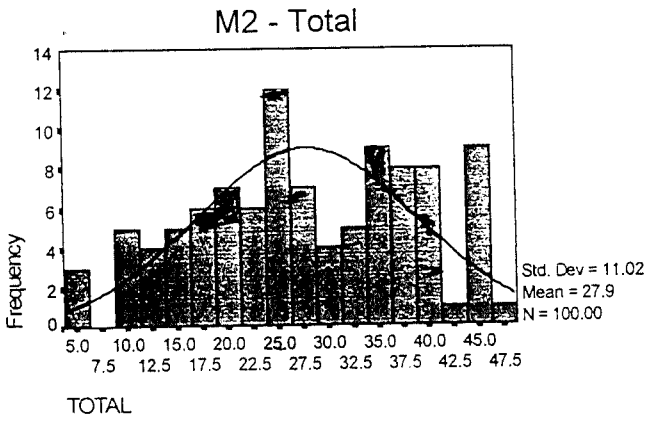

M1 - Total
Frequency
Std. Dev = 9.55
Mean = 19.6
N = 100.00
TOTAL

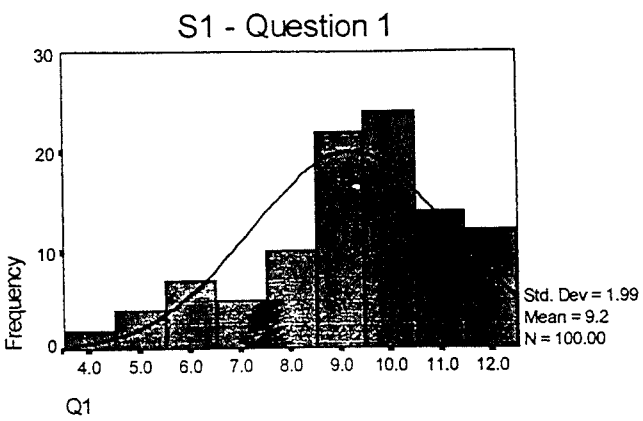

M2 - Question 1
Frequency
Std. Dev = 3.44
Mean = 6.7
N = 100.00
Q1

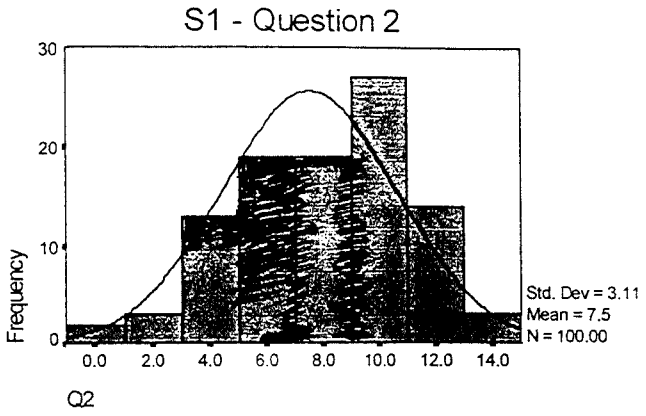

M2 - Question 2
Frequency
Std. Dev = 3.50
Mean = 7.5
N = 100.00
Q2

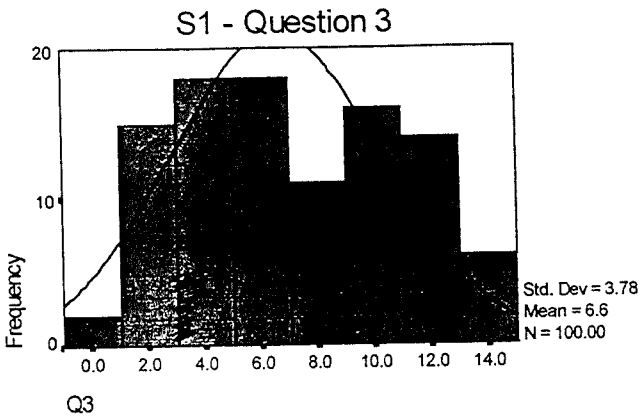

M2 - Question 3
Frequency
Std. Dev = 3.36
Mean = 6.4
N = 100.00
Q3

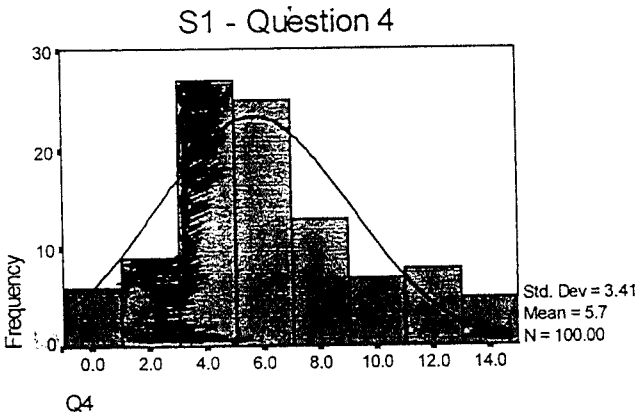

M2 - Question 4
Frequency
Std. Dev = 4.07
Mean = 7.2
N = 100.00
Q4

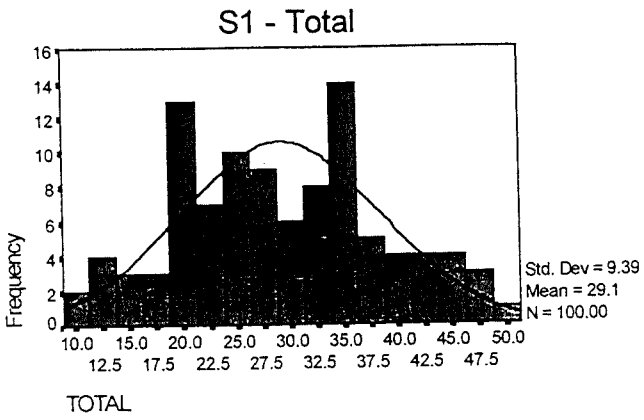

M2 - Total
Frequency
Std. Dev = 11.02
Mean = 27.9
N = 100.00
TOTAL

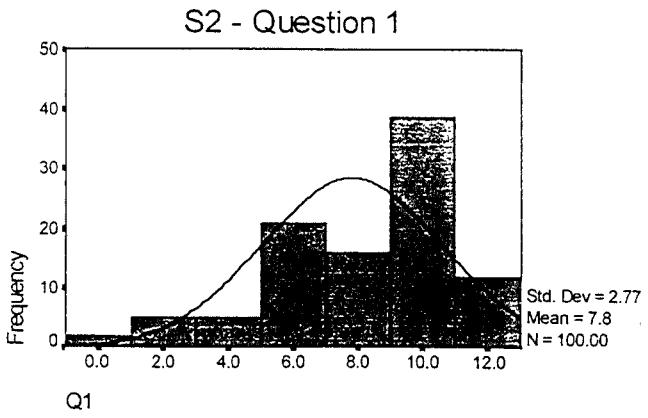

S1 - Question 1
Frequency
Std. Dev = 1.99
Mean = 9.2
N = 100.00
Q1

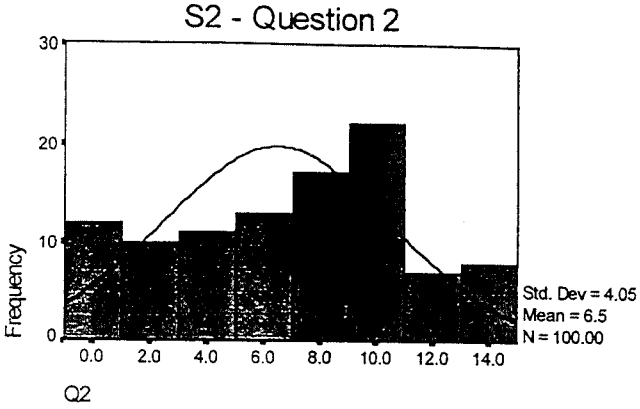

S1 - Question 2
Frequency
Std. Dev = 3.11
Mean = 7.5
N = 100.00
Q2

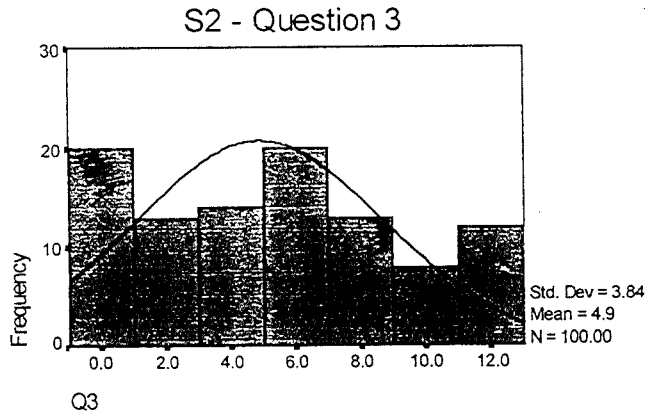

S1 - Question 3
Frequency
Std. Dev = 3.78
Mean = 6.6
N = 100.00
Q3

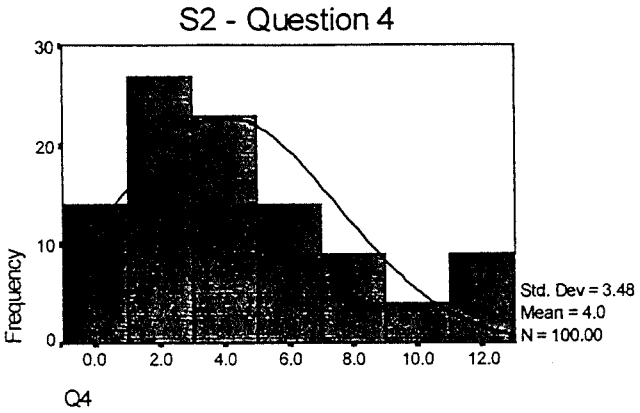

S1 - Question 4
Frequency
Std. Dev = 3.41
Mean = 5.7
N = 100.00
Q4

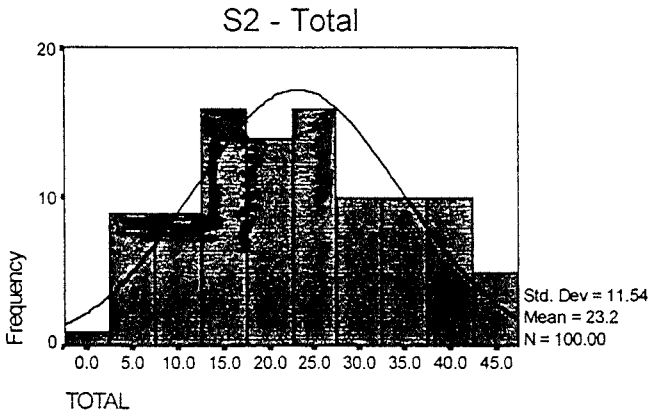

S1 - Total
Frequency
Std. Dev = 9.39
Mean = 29.1
N = 100.00
TOTAL


S2 - Question 1
Frequency
Std. Dev = 2.77
Mean = 7.8
N = 100.00
Q1


S2 - Question 2
Frequency
Std. Dev = 4.05
Mean = 6.5
N = 100.00
Q2


S2 - Question 3
Frequency
Std. Dev = 3.84
Mean = 4.9
N = 100.00
Q3


S2 - Question 4
Frequency
Std. Dev = 3.48
Mean = 4.0
N = 100.00
Q4


S2 - Total
Frequency
Std. Dev = 11.54
Mean = 23.2
N = 100.00
TOTAL

Likelihood Ratios

The following are the likelihood ratios from the models described in chapter 6

Module	-2*log(lh)
1	33164.3
2	36301.4
3	37182.4
4	1754.6
5	522.4
6	147.3
7	26263.5
8	16057.2
9	3878.8
13	30799.0
14	19296.9
15	7745.1
19	2819.5
21	242.3
Combined 3 level	208401.0

The likelihood ratios for the modelling of chapter 8 are

Year	-2*log(lh)
1994	8272.1
1995	12061.2
Combined	20454.5

BIBLIOGRAPHY

Adams R and Wilmut J (1981) *A Measure of the Weights of Examination Components, and Scaling to Adjust Them;* The Statistician 30/ 4

Angoff, W. H. (1984) *Scales, Norms and Equivalent Scores;* Princetown, New Jersey; Educational Testing Service

Arnot M, David M and Weiner G (1996) *Educational Reforms and Gender Equality in Schools;* Manchester; Equal Opportunities Commission

Backhouse J K (1976) *Determination of Grades for two Groups Sharing a Common Paper;* Educational Research Volume 18, p126-137

Bardell G S, Forrest G M and Shoesmith D J (1978) *Comparability in GCE* JMB

Beecher T and Maclure S (1978) *The Politics of Curriculum Change;* London; Hutchinson & Co.

Bell A W, Costello J and Küchemann (1983) Research on Learning and Teaching - A Review of Research in Mathematical Education p254; Windsor NFER-Nelson Publishing Co Ltd.

Beloe R (1960) *Secondary School Examinations other than GCE;* London; HMSO

Bennett N and Dunne E (1994) *Managing Groupwork;* In Moon B and Mayes A S eds. (1994) *Teaching and Learning in the Secondary School; p166-172;* London; Routledge

Biggs J B and Collis A F (1982) *Evaluating the Quality of Learning: The SOLO Taxonomy;* New York; Academic Press

Bloom B S (1956) *Taxonomy of Educational Objectives. Handbook 1: The Cognitive Domain;* New York; Mckay

Bloomfield B, Dobby J and Duckworth D (1977) *Mode Comparability in the CSE* Schools Council Examinations Bulletin 36 Evans/Methuan Educational

Bloomfield B, Dobby J and Kendall L (1979) *Ability and Examinations at 16+;* London; Macmillan Education

Brereton J L (1944) *The Case for Examinations;* Cambridge; Cambridge University Press

Brereton J L (1965) *Exams: Where Next?;* Victoria, Canada; Pacific Northwest Humanist Publications

Broadfoot P (1988) *Records of Achievement and the National Assessment Framework;* Paper presented to a one day BERA conference 11.2.88

Broadfoot P ed. (1984) *Selection, Certification and Control;* Lewes, Sussex; Falmer Press

Broadfoot P (1996) *Education, Assessment and Society;* Buckingham; Open University Press

Brualdi Amy (1999) *Traditional and Modern Concepts of Validity;* Washington DC; ERIC Clearinghouse on Assessment and Evaluation

Christie T and Forrest G M (1981) *Defining Public Examination Standards* Schools Council

Cockcroft W H (1982) *Mathermatics Counts;* London; HMSO

CVCP (1962) *A |Report of a Sub-Committee on University Entrance Requirements in England and Wales;* London; Association of Universities of the British Commonwealth

Cresswell M (1993) *Public Examination Statistics as Indicators of Comparability* Paper presented at the IGRC seminar on the Interpretation of Examination Statistics SEG

Cresswell M (1996) *Defining, Setting and Maintaining Standards in Curriculum-embedded Examinations:* In Goldstein H and Lewis T eds (1996) *Judgemental and Statistical Approaches; Assessment: Problems, Developments and Statistical Issues; p57-84;* John Wiley & Sons Ltd.

Cronbach L J (1951) *Coefficient Alpha and the Internal Structure of Tests* Psychometrica 16/3

Cronbach L J and Meehl P E (1955) *Construct Validity in Psychological Tests;* Psychological Bulletin 52/4

Deale R N (1975) *Assessment and Testing in the Secondary School* Evans/Methuen Educational

Dearing R (1995) *Review of 16-19 Qualifications: Interim Report;*

Dearing R (1996) *Review of Qualifications for 16-19 Year Olds;* Hayes; SCAA Publications

Dewey J (1930) *Democracy in Education;* New York; The Macmillan Company

Ebel R L (1965) *Measuring Educational Achievement* Prentice-Hall

Expenditure Committee of the House of Commons (1977) *The Attainments of the School Leaver;* London; HMSO

Forrest G M and Shoesmith D J (1985) *A Second Review of Comparability Studies;* Manchester; JMB

Fowles D (1974) *CSE: two research studies;* London; Evans/Methuan Educational

French S, Slater J B, Vassiloglou M and Willmott A S (1987) *Descriptive and Normative Techniques in Examination Assessment;* Oxford; University of Oxford Delegacy of Local Examinations

Gardner H (1983) *Frames of Mind;* New York; Basic Books

GCE Boards of England, Wales and Northern Ireland (1983) *Common Cores at Advanced Level*

GCE Examining Boards of England, Wales and Northern Ireland (1993) *Criteria and Procedures for the Implementation of the Principles for GCE Advanced Level and Advanced Supplementary Examinations*

Gipps C (1994) *Beyond Testing;* London; The Falmer Press

Glaser R (1963) *Instructional Technology and the Measurement of Learning Outcomes* American Psychologist 18

Glass G V (1978) *Standards and Criteria* Journal of Educational Measurement 15/4

Goldschlager L and Lister A *Computer Science A Modern Introduction p74;* London; Prentice Hall International Inc.

Goldstein H (1987) *Multilevel Models in Educational and Social Research* Griffin

Goldstein H (1995) *Multilevel Statistical Models;* London; Arnold (Second Revised Edition of above)

Goldstein and Cresswell M J (1996) *The Comparability of Different Subjects in Public Examinations: A Theoretical and Practical Critique;* Oxford Review

Goldstein H and Thomas S (1995) *Questionable Value;* Education 185/11

Good F

Good F and Cresswell M (1988) *Grading the GCSE;* Secondary Examinations Council;

Good F and Cresswell M (1988) *Differentiated Assessment: Grading and Related Issues: Vol 1;* SEC

Gray E A (1992) *MEI Questionnaire;* Unpublished

Gray J, Jesson D, Goldstein H, Hedger K and Rasbash J (1995) *A Multi-Level Analysismof School Improvement: Changes in School Performance over Time;* School Effectiveness and School Improvement 6/1

Guilford J and Fruchter B (1981) *Fundamental Statistics in Psychology and Education;* McGraw-Hill Inc.

Hall C G *The Structure of Some Educational Abilities at the 16+ A Level in England and Wales;* D. Phil thesis, University of Brunel

Hargreaves D H (1982) *The Challenge for the Comprehensive School;* London; Routledge & Kegan Paul Ltd.

Harlen W, Gipps C, Broadfoot P and Nuttall D (1994) *Assessment and the Improvement of Education;* In Moon B and Mayes A S eds. (1994) *Teaching and Learning in the Secondary School p273-286;* London; Routledge

Hart J (1988) *The Scottish Action Plan Experience;* in Moon B ed. (1988) *Modular Curriculum p107-140;* London; PCP Ltd.

Higginson G (1988) *Advancing A Levels;* London; HMSO

Howat G M D (1974) *OCSEB 1873 - 1973;* Oxford; OCSEB.

Hughes S (1995) *What Makes GCSE Questions Difficult;* Cambridge; UCLES

Ingenkamp Karlheinz (1977) *Educational Assessment;* Windsor; NFER Publishing Company

Jenkins and Walker eds. (1994) *Developing Student Capability through Modular Courses;*

Johnson S and Cohen L (1983) *Investigating Grade Comparability through Cross-Moderation* Schools Council

Kingdon M and Stobart G (1988) *GCSE Examined;* London; The Falmer Press

Kingdon M (1991) *The Reform of Advanced Level;* London; Hodder and Stoughton

Lindquist E F ed. (1950) *Educational Measurement;* Washington; American Council on Education

Lord F and Novick M (1968) *Statistical Theories of Mental Test Scores;* Massachusetts; Addison-Wesley Publishing Company

Matthews J C and Leece J R (1976) *Examinations:Their Use in Curriculum Evaluation and Development; SC Examinations Bulletin 33;* London; Evans/Methuan Educational

Messick S (1989) *Validity* in **Linn R ed.** *Educational Measurement;* Washington; American Council on Education/Macmillan

Moll L (1990) *Vygotsky and Education;* Cambridge; Cambridge University Press

Moon B (1986) *The New Maths Curriculum Controversy;* Lewes; The Falmer Press

Moon B ed. (1988) *Modular Curriculum;* London; PCP Ltd.

Murphy R and Torrance H (1988) *The Changing Face of Educational Assessment;* Milton Keynes; Open University Press

Nickson M (1994) *Evaluation of the Module Bank System;* Cambridge; UCLES

Newbould C and Massey A (1979) *Comparability using a Common Element;* Cambridge; TDRU

Newton P (1997) *Measuring Comparability of Standards between Subjects:Why Our Statistical Techniques Do Not Make the Grade;* British Educational Research Journal 23/4

Norwood (1941) *Curriculum and Examinations in Secondary Schools;* London; HMSO

Nuttall D L and Willmott A (1972) *Techniques of Analysis;* Slough; NFER Publishing Co.

Nuttall D L, Backhouse J K and Willmott A S (1974) *Comparability of Standards between Subjects;* Evans/Methuen Educational

Nuttall D L (1987) *The Validity of Assessments;* European Journal of Psychology of Education, vol 2 no 2 p109-118

O'Donoghue C, Thomas S, Goldstein H and Knight T (1996) *1996 DfEE Study of Value Added for 16-18 Year Olds in England;*

Orr L and Nuttall D L (1983) *Comparability in Examinations: Occasional Paper 2: Determining standards in the proposed single system of examining at 16+;* Schools Council

Papert S (1980) *Mindstorms: Children, Computers and Powerful Ideas;* Brighton; The Harvester Press

Pearce J (1972) *School Examinations;* London; Collier-Macmillan

Pollitt A (1993) *Aligning Standards: psychometric contributions to examination problems;* Paper presented to the Aligning Standards Seminar, UCLES

Pollitt A, Hutchinson C, Entwhistle N and de Luca C (1985) *What Makes Exam Questions Difficult?;* Research Reports for Teachers 2 University of Edinburgh

Pollitt A, Hughes S, Ahmed A, Fisher-Hoch H and Bramley T (1998) *The Effect of Structure on the Demands in GCSE and A Level Questions;* (Unpublished)

Popper K (1972) *Objective Knowledge: An Evolutionary Approach;* Oxford; Oxford University Press

Porkess K (1995) *MEI Structured Mathematics: Five Years On;* Bradford on Avon; MEI

Quinlan M (1993) *A Comparability Study in GCSE History;* ULEAC

Quinlan M (1997) *A Comparability Study in A level Mathematics;* Unpublished

Ratcliffe P (1993) *A Comparability Study in GCSE Geography;* NEAB

Reid W A (1972) *The Universities and the Sixth Form Curriculum;* Basingstoke; Macmillan Education Ltd.

Rust J and Golombok S (1989) *The Science of Psychological Assessment;* London; Routledge

SCAA (1993, revised 1996) *GCE A and AS Code of Practice;* London; SCAA Publications

SCAA (1996) *Evaluation of Modular A Levels;* SCAA\140558\1

Schools Council Forum on Comparability (1979) *Comparability in Examinations : Occasional Paper 1: Standards in public examinations: problems and possibilities* Schools Council

Schools Council Working Paper 45 (1972) *16 - 19 Growth and Response;* Surrey; Evans/Methuan Educational

Scottish Education Department (1980) *The Munn and Dunning Reports: The Government's Development Programme;* SED

SEAC (1989) *Consultation on the Secretaries of State's Remit to the School Examinations and Assessment Council on the Promotion of AS Examinations and the Rationalisation of the A Level Syllabuses;* London; SEAC

SEC (1987) *Assessing Modular Syllabuses: A Discussion Document;* SEC

Siegel H (1988) *Educating Reason: Rationality, Critical Thinking and Education;* New York; Routledge

SRAC (1990) *Standards in Advanced Level Mathematics Report of Study 1 A Study of the Demands made by the Two Approaches to "Double Mathematics"* UCLES

SRAC (1990) *Standards in Advanced Level Mathematics Report of Study 2 Survey of Resources for Double-Subject Mathematics* OCSEB

SRAC (1990) *Standards in Advanced Level Mathematics Report of Study 3 The Boards' Approaches to Double-Subject Mathematics* AEB

SRAC (1990) *Standards in Advanced Level Mathematics Report of Study 4 A Statistical Study of Double-Subject Mathematics Candidates at the June 1986 Examinations* AEB

SRAC (1990) *Standards in Advanced Level Mathematics: Report of Study 5; A Cross-Moderation Exercise based on Scripts from the June 1987 Examinations in Advanced Level Mathematics and Report of Study 6; A Statistical Survey based on the June 1987 Examinations in Advanced Level Mathematics* NISEAC

SRAC (1990) *A and AS Modular Syllabuses and their Assessment: a Report to the GCE Secretaries*

Stobart G, Elwood J, Hayden M, White J & Mason L (1992) *Differential Performance in Examinations at 16+: English and Mathematics;* ULEAC/NFER

Taverner S and Wright M (1995) *A Review of Modular A Level Mathematics;* University of Newcastle Department of Education

Theodossin E (1986) *The Modular Market;* Further Education Staff College

Thomson D G (1992) *Grading Modular Curricula;* Cambridge; Midland Examining Group

Thorndike, R. L. & Hagen, E (1961) *Measurement and Evaluation in Psychology and Education, Second Edition;* New York; John Wiley and Sons

Tymms P B and Fitz-Gibbon C T (1991) *A Comparison of Examining Boards A Levels* Oxford Review of Education 17/1

Tymms P B and L Vincent (1994) *Comparing Examination Boards and Syllabuses at A Level: students' grades, attitudes and perceptions of classroom processes;* CEM centre Department of Education University of Newcastle upon Tyne

UCAS (1994) *Examinations and Grades;* Cheltenham; UCAS

Wiliam D (1993) *Reconceptualising Validity, Dependability and Reliability for National Curriculum Assessment;* BERA Conference, 1993

Wiliam D and Black P (1995) *Meanings and consequences: a basis for distinguishing formative and summative functions of assessment?* Kings College London; BERA Conference, 1995

Willmott A and Hall C (1975) *O-Level Examined: The Effects of Question Choice;* London; Schools Council

Wood R (1991) *Assessment and Testing;* Cambridge; Cambridge University Press

GLOSSARY

ACAC	Curriculum and Assessment Authority for Wales
ALIS	A Level Information Service
AEB	Associated Examining Board
CEM	Curriculum, Evaluation and Management Centre, University of Newcastle upon Tyne
CEE	Certificate of Extended Education (17+)
CNAA	Council for National Academic Awards
CSE	Certificate of Secondary Education (16+)
CVCP	Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom
ERA	Education Reform Act (1988)
DFE	Department for Education
DES	Department of Education and Science
FE	Further Education
GCE	General Certificate of Education O - Ordinary level (16+) A - Advanced Level (18+) AS - Advanced Supplementary (17/18+)
GCSE	General Certificate of Secondary Education (16+)
HMC	Headmasters' Conference
IB	International Baccalaureat
LEA	Local Education Authority
Mode 1	An examination whose syllabus, examination and assessment are conducted by the accrediting board
Mode 2	An examination whose syllabus is provided by the school (or group of schools), but for which the examination and assessment are conducted by the board
Mode 3	An examination for which the school (or school group) who provide the syllabus, also set and mark the examination
MEI	Mathematics in Education and Industry
NEAB	Northern Examinations and Assessment Board
NUJMB	Northern Universities Joint Matriculation Board
NCC	National Curriculum Council (1988-1993)
NCVQ	National Council for Vocational Qualifications
NFER	National Foundation for Educational Research
OCEAC	Oxford and Cambridge Examinations and Assessment Council
OCSEB/O&C	Oxford and Cambridge Schools Examination Board
OED	Oxford English Dictionary
OU	Open University
QCA	Qualifications and Curriculum Authority (1997-)
ROSLA	Raising of the School Leaving Age (1972)
SAP	Scottish Action Plan
SAT	Standard Assessment Task (UK) Scholastic Aptitude Test (US)

SCAA	School Curriculum and Assessment Authority (1993-1997)
SC	Schools Council (1964-1982)
SSCC	Secondary Schools Curriculum Council
SCDC	Schools Curriculum Development Committee (1982-1988)
SEAC	Schools Examinations and Assessment Council (1988-1993)
SEC	Secondary Examinations Council (1982-1988)
SMP	Schools Mathematics Project
SRAC	Standing Research Advisory Commitee
SSEC	Secondary Schools Examination Council (1917-1963)
TGAT	Task Group on Assessment and Testing
TVEI	Technical and Vocational Education Initiative
UCAS	Universities and Colleges Admissions Service
UCL	Universlty College, London University
UCLES	University of Cambridge Local Examinations Syndicate[1]
ULEAC	University of London Examinations and Assessment Council (now EDEXEL)
UODLE	University of Oxford Delegacy of Local Examinations
WJEC	Welsh Joint Examinations Committee

[1] The A level suites of UCLES, OCSEB and UODLE are now administered by OCR.